AD________________

Award Number: W81XWH-07-1-0503

TITLE: Identification of Novel Genes and Candidate Targets in CML Stem Cells

PRINCIPAL INVESTIGATOR: Dr. Connie Eaves

CONTRACTING ORGANIZATION: British Columbia Cancer Agency
Vancouver BC, V5Z 1L3

REPORT DATE: January 2009

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE 1 Jan 2009	**2. REPORT TYPE** Final	**3. DATES COVERED** 1 Jul 2007 – 31 Dec 2008

4. TITLE AND SUBTITLE

Identification of Novel Genes and Candidate Targets in CML Stem Cells

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-1-0503

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dr. Connie Eaves
Dr. Yun Zhao

E-Mail: ceaves@bccrc.ca

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

British Columbia Cancer Agency
Vancouver BC, V5Z 1L3

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Chronic myeloid leukemia (CML) is believed to originate from a normal hematopoietc stem cell acquiring the BCR-ABL fusion gene whose protein product has hyperactive tyrosine kinase activity. Though imatinib mesylate(IM) that targets BCR-ABL kinase activity is now widely used, its curative potential as a single agent is not sure, moreover it is unlikely to eliminate the CML stem cells either, which highlights the necessity to elucidate the molecular mechanism operative in CML stem cells. Previously 16 LongSAGE libraries were established to analyze the CML stem cell and their normal counterparts from various sources. There are numerous novel tags which might represent bona fidetranscripts uniquely expressed in these primitive CML SAGE libraries, which provide us an uniquely opportunity to discovery unknown but important transcripts in these cells. We utilized bioinformatics analyses to sort out these novel tags as the candidates to recover the potential bona fide transcripts, and then with PCR and 3'-RACE (Rapid Amplification of cDNA End) approaches we assessed the validity of them and recovered the 3'-end of the potential novel transcripts originated from these tags with the sequence confirmation. The 5'-RACE is under way to eventually recovery the full-length cDNAs and their gene expression pattern between multiple CML and normal primitive cell samples will be assessed.

15. SUBJECT TERMS
CML stem cells, Global transcriptome analysis, Bioinformatic manipulation of LongSAGE data

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	32	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

	Page
Introduction	4
Body	4
Key Research Accomplishments	5
Reportable Outcomes	6
Conclusion	6
References	6
Supporting Data	8
Appendices	17

INTRODUCTION:

Chronic myeloid leukemia (CML) is a blood malignancy that is believed to originate in a hematopoietic stem cell as a result of the formation of an abnormal fusion gene called *BCR-ABL*, whose protein product is a deregulated tyrosine kinase[1-3]. The CML stem cells derived from the progeny of this first affected cell produce deregulated numbers of various mature blood cells as well as more abnormal stem cells that allow the disease to be perpetuated and further evolve. The CML stem cells are also likely responsible for some CML patients' resistance to imatinib mesylate (IM), a recently developed anti-ABL drug now used as front line treatment for CML patients worldwide[4,5]. Because of the very low frequency of CML stem cells in most patients with CML, an understanding of the molecular make-up and regulation of these cells has been difficult to obtain[6-8]. Recently, we developed and validated a "PCR-LongSAGE" methodology that makes it possible to produce LongSAGE libraries from as few as 1000 cells. We then used this approach to profile very primitive CML cells from selected patients with predominantly Ph+ stem cells and also the counterparts of these subsets obtained from normal individuals[9]. A unique feature of the LongSAGE approach, is that it enables the discovery and identification of novel transcripts originating from novel tags[9-11]. Such tags unique to primitive CML cells would be of particular interest as they could be anticipated to encode unknown proteins or non-coding RNAs of disease relevance. They are also likely to provide new clues to the designing of more effective therapies against CML. To pursue this overall goal, we first identified a large number of novel tags uniquely present in the libraries we had already prepared from primitive CML cells. The tags were then interrogated with multiple bioinformatics approaches suitable for full-length cDNA recovery. A PCR approach was utilized either to confirm the validity of these tags, or to recover and sequence the 3'-end and 5'end of the novel transcripts.

BODY:

I. Use of a PCR approach to confirm the validity of novel tags identified in a human bone marrow lin$^-$ CD34$^+$ cell library

We generated a LongSAGE library from normal adult human bone marrow (BM) lin$^-$CD34$^+$ cell RNA and sequenced it to a depth of 200,000 tags. We used DiscoverySpace software[12] to compare our library with 2 short SAGE libraries generated by others from similar starting cell populations[13,14] and found our library to be 96% and 94% similar to these. We also compared our library to 287 publicly accessible SAGE libraries prepared from multiple types of human cells (available primarily through CGAP(http://cgap.nci.nih.gov) as well as RefSeq (ftp://ftp.ncbi.nih.gov/refseq/daily), MGC (ftp://ftp.ncbi.nih.gov/repository/MGC/MGC.sequences), Ensembl and EST databases. This allowed us to identify 23 tags that map to a single conserved (in mouse and rat) site in the human genome that are unique in our lin$^-$CD34$^+$ LongSAGE BM library. These 23 novel tags are listed with their chromosomal locations in Supporting data - Table 1. Q-RT-PCR was then used to investigate the expression of these 23 novel tags in 3 different lin$^-$CD34$^+$ adult human BM cell cDNA preparations, including one from the same pool of RNA used for making the original SAGE library. To assess the possibility of genomic DNA contamination and its contribution to the detection of the unique tag expression, we included a strict negative control in which RNA from each BM sample was used as a PCR template. Q-RT-PCR analyses showed 10 of the 23 tags to be consistently detectable in the cDNA samples examined with no detectable amplification in the negative controls[9].

II. 3'- and 5'- RACE recovery and sequence verification of novel tags expressed uniquely in primitive CML cells

We first created a meta-library from 6 CML libraries (~1.2 million total tags, 171,000 different tag types) constructed from the CD38$^+$ and CD38$^-$ subsets of lin$^-$CD34$^+$ leukemic (Ph+) cells obtained from 3 chronic phase CML patients. Over 109,000 unique tags have been identified by comparing this meta-library with another meta-library that included 10 LongSAGE libraries including corresponding subsets of primitive cells from normal adult human BM, G-CSF-mobilized peripheral blood, cord blood and fetal liver (~2 million total tags, 224,000 different tag types)[9,15]. The detailed information of these libraries is summarized in Supporting data - Table 2. These unique tags have been further selected by removal of tags present in the human CGAP pool of 42 LongSAGE (21-mer) libraries (which contain ~500,000 different tag types) to yield 89,730 unique tags. We then converted these long tags to short tags (which yielded 72,614 tags) and removed any tags present in the human CGAP pool of 272 short SAGE (14-mer) libraries (which contain ~621,000 different tag types). The 6824 short tags thus obtained were converted back to their original long tags (7067) which were then filtered bioinformatically to yield 2 categories of novel tags. 1) The first category consisted of tags that had a single site in the human genome, no overlap with any known cDNA (identified in RefSeq, MGC or Ensembl) or EST database; appeared at least twice in all the human LongSAGE libraries generated at our centre and were >2kb away from any known transcript in the human genome (Figure 1). 69 tags that fulfilled these criteria were unique to the CML lin$^-$CD34$^+$CD38$^{-/+}$ metalibrary (Supporting data - Table 3). 2) To identify novel tags in the CML lin$^-$CD34$^+$CD38$^-$ metalibrary, we further excluded any tag present in 3 lin$^-$CD34$^+$ MDS cell and 4 LongSAGE libraries prepared from different primitive and mature subsets of human mammary gland cells[16]. This yielded 39 tags in this second category (Table 4).

For tags unique to CML CD34$^+$ cells, we used RLM-RACE technology (Ambion, Applied Biosystems) to clone potential novel transcripts from a sample of RNA pooled from several human myeloid leukaemia cell lines (K562, HL60, KG1 and TF1). We then extracted the 5' genomic sequence of each individual tag (Genome Browser, UCSC) and used Integrated DNA Technology (http://www.idtdna.com/SciTools) online software to design 2 primers for each tag located <100 base pairs away from the tag. We performed nested 3'-RACE for the 69 novel tags and 21 showed positive nested PCR products (Supporting data - Figure 2a), with an average size of 350 bp (ranging from 200-800 bp). For 5 of these 21 tags, multiple 3'RACE products were obtained. 10 out of these 21 were continued to perform 5'-RACE experiments and showed the positive nested PCR results. The PCR products from tags having both successful 3'- and 5'- RACE were subcloned into TOPO vectors and were sequenced. Eventually, cDNA of both 3'- and 5'- RACE derived from 5 tags were fully sequence verified. In the 3'-RACE experiment, one case of an alternative 3'-end of cDNA, which showed a 133 bp difference between the longer 3'-end and the shorter form, mentioned above was sequence confirmed. (3'-RACE products generated from tag#003 shown in the Supporting data - Table 3 and the RACE result shown in Supporting data – Figure 2a.).

Analysis of the fragments from the *bona fide* transcripts using BLAT through Genome Browser, showed that #33 and #46 could represent alternative 3'-UTRs (Untranslated Regions). #33 and # 23 could represent novel transcripts in either an isolated genomic region or in an opposite direction though close to a known transcript; #32 represents a new transcript though some ESTs are nearby (Supporting data Figure 3). We then designed new primers either to test the validity of the transcript or to quantify its level of expression (Supporting data Figure 3 & 4). The cDNA fragment recovered from #033 tag is a spliced sequence as well, thus it's a suitable candidate for quantitative measurement of its expression in different primary CML and BM samples. 9 CML and 2 BM lin$^-$CD34$^+$ cells were compared and the result showed the tag-derived cDNA indicated a more than 2-fold higher expression in CML cells than in BM cells although not this difference did not achieve significance ($p = 0.1$, Supporting data Figure 4). Considering that this novel tag was a singleton present in only one CML library, the Q-RT-PCT measurement is encouraging. The particular gene affected is KIF5B, which was found as a rare partner fused with PDGFA in a patient with idiopathic hypereosinophilic syndrome (IHES) which disappeared upon IM. Interestingly, it is known that IM can also target PDGFA [17]. The alternative UTR we identified here suggested complicated translational regulation of this gene and its potential functional role in either normal or leukemic hematopoiesis will undoubtedly draw more attention.

To assess the validity of these unique tags from the CML CD34$^+$CD38$^-$ cells, we looked up the 5' and 3' genomic sequence of each individual tag using Genome Browser of UCSC and then used IDT (Integrated DNA Technology, http://www.idtdna.com/SciTools) online software to design 2 primers on each side of the tag located <100 base pairs away from the tag for use in an RT-PCR study. New RNA extracts were obtained from CD34$^+$CD38$^-$ CML cells (including from one of the samples used to construct the original LongSAGE libraries) and cDNA prepared from each. RNA from the same samples was run as a control in the PCR analyses subsequently performed. These detected cDNA-specific PCR products of the expected size for 3 of the 31 tags tested (Supporting data - Figure 2b). This approach was described in part I and was used elsewhere to demonstrate the validity of novel tags[9,10].

KEY RESEARCH ACCOMPLISHMENTS:

I. In silico discovery of 23 novel tags found unique to the normal adult bone marrow lin$^-$CD34$^+$ LongSAGE library we generated, 10 of which could be validated by RT-PCR.

II. In silico discovery of 69 novel tags unique to CML lin$^-$CD34$^+$ cells, 5 of which could be recovered using both 3'- and 5'-RACE and verified with various PCR and sequencing experiments. One tag derived cDNA was measured with Q-RT-PCR using primary CML and BM cells as well and showed higher expression level in the CML cells although the difference was not statistically significant.

III. In silico discovery of 39 novel tags unique to CML lin$^-$CD34$^+$CD38$^-$ cells, 3 of which could be detected by RT-PCR analysis of extracts of lin$^-$CD34$^+$CD38$^-$ cells from 3/3 different chronic phase CML patients.

REPORTABLE OUTCOMES:

Zhao Y, Raouf A, Kent D, Khattra J, Delaney A, Schnerch A, Asano J, McDonald H, Chan C, Jones S, Marra MA, Eaves CJ. A modified polymerase chain reaction-long serial analysis of gene expression protocol identifies novel transcripts in human CD34+ bone marrow cells. Stem Cells. 25:1681-9, 2007.

Salvagiotto G, Zhao Y, Vodyanik M, Ruotti V, Stewart R, Marra M, Thomson J, Eaves C, Slukvin I. Molecular profiling reveals similarities and differences between primitive subsets of hematopoietic cells generated in vitro from human embryonic stem cells and in vivo during embryogenesis. Exp Hematol, 36:1377-89, 2008

Zhao Y, Delaney A, Marra M, Eaves AC, Eaves CJ. Comparative transcriptome analysis of normal and chronic myeloid leukemia stem cells. Exp Hematol. 35 (Suppl. 2):61, 2007.

Zhao Y, Delaney A, Marra M, Jiang X, Eaves AC, Eaves CJ. Comparative transcriptome analysis of different subsets of CD34+ normal and chronic myeloid leukemia cells identifies novel perturbations in the CML stem cell population. Blood. 110 (Suppl. 1):19a, 2007.

Zhao Y, Delaney A, Raouf A, Raghuram K, Li HI, Schnerch A, Jiang X, Eaves AC, Marra MA, & Eaves CJ. Differentially expressed and novel transcripts in highly purified chronic phase CML stem cells. Blood. 112 (Suppl. 1): 79a, 2008.

CONCLUSION:

We used various approaches to isolate, validate and recover novel tags uniquely expressed in subsets of primitive CML cells. Together, these findings indicate that the novel tags could represent *bona fide* transcripts. Since these newly discovered transcripts are uniquely expressed in rare primitive CML cells, they may be in the maintenance of the disease or its resistance to available therapies.

These studies set the stage for further full-length cDNA recovery, gene structure analysis, and gene expression assessment between primitive CML and normal cells. Final functional analyses may provide a unique opportunity to obtain novel clues for designing improved therapies.

REFERENCES:

1. Rowley JD. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 243:290-293, 1973.
2. Konopka JB & Witte ON. Detection of c-*abl* tyrosine kinase activity in vitro permits direct comparison of normal and altered *abl* gene products. Mol Cell Biol 5:3116-3123, 1985.
3. Ben-Neriah Y, Daley GQ, Mes-Masson AM, Witte ON & Baltimore D. The chronic myelogenous leukemia-specific P210 protein is the product of the bcr/abl hybrid gene. Science 233:212-214, 1986.
4. Copland M, Hamilton A, Elrick LJ, Baird JW, Allan EK, Jordanides N, Barow M, Mountford JC & Holyoake TL. Dasatinib (BMS-354825) targets an earlier progenitor population than imatinib in primary CML, but does not eliminate the quiescent fraction. Blood 107:4532-4539, 2006.
5. Jiang X, Saw KM, Eaves A & Eaves C. Instability of BCR-ABL gene in primary and cultured chronic myeloid leukemia stem cells. J Natl Cancer Inst 99:680-693, 2007.
6. Ohmine K, Ota J, Ueda M, Ueno S, Yoshida K, Yamashita Y, Kirito K, Imagawa S, Nakamura Y, Saito K, Akutsu M, Mitani K, Kano Y, Komatsu N, Ozawa K & Mano H. Characterization of stage progression in chronic myeloid leukemia by DNA microarray with purified hematopoietic stem cells. Oncogene 20:8249-8257, 2001.
7. Kronenwett R, Butterweck U, Steidl U, Kliszewski S, Neumann F, Bork S, Blanco ED, Roes N, Graf T, Brors B, Eils R, Maercker C, Kobbe G, Gattermann N & Haas R. Distinct molecular phenotype of malignant CD34$^+$ hematopoietic stem and progenitor cells in chronic myelogenous leukemia. Oncogene 24:5313-5324, 2005.

8. Diaz-Blanco E, Bruns I, Neumann F, Fischer JC, Graef T, Rosskopf M, Brors B, Pechtel S, Bork S, Koch A, Baer A, Rohr UP, Kobbe G, Haeseler A, Gattermann N, Haas R & Kronenwett R. Molecular signature of CD34$^+$ hematopoietic stem and progenitor cells of patients with CML in chronic phase. Leukemia 21:494-504, 2007.

9. Zhao Y, Raouf A, Kent D, Khattra J, Delaney A, Schnerch A, Asano J, MacDonald H, Chan C, Jones S, Marra MA & Eaves CJ. A modified polymerase chain reaction-long serial analysis of gene expression protocol identifies novel transcripts in human CD34$^+$ bone marrow cells. Stem Cells 25:1681-1689, 2007.

10. Siddiqui AS, Khattra J, Delaney A, Zhao Y, Astell C, Asano J, Babakaiff R, Barber S, Beland J, Bohacec S, Brown-John M, Chand S, Charest D, Charters AM, Cullum R, Dhalla N, Featherstone R, Gerhard DS, Hoffman B, Holt R, Hou J, Kuo BY-L, Lee LLC, Lee S, Leung D, Ma K, Matsuo C, Mayo M, McDonald H, Prabhu A-L, Pandoh P, Riggins GJ, de Algara TR, Rupert JL, Smailus D, Stott J, Tsai M, Varhol R, Vrljicak P, Wong D, Wu MK, Xie Y-Y, Yang G, Zhang I, Hirst M, Jones SJM, Helgason CD, Simpson EM, Hoodless PA & Marra M. A mouse atlas of gene expression: Large-scale, digital gene expression profiling resource from precisely defined developing C57BL/6J mouse tissue and cells. Proc Natl Acad Sci USA 102:18485-18490, 2005.

11. Hirst M, Delaney A, Rogers SA, Schnerch A, Persaud DR, O'Connor MD, Zeng T, Moksa M, Fichter K, Mah D, Go A, Morin RD, Baross A, Zhao Y, Khattra J, Prabhu A-L, Pandoh P, McDonald H, Asano J, Dhalla N, Ma K, Lee S, Ally A, Chahal N, Menzies S, Siddiqui A, Holt R, Jones S, Gerhard DS, Thomson JA, Eaves CJ & Marra MA. LongSAGE profiling of nine human embryonic stem cell lines. Genome Biol 8:R113 2007.

12. Robertson N, Oveisi-Fordorei M, Zuyderduyn SD, Varhol RJ, Fjell C, Marra M, Jones S & Siddiqui A. DiscoverySpace: an interactive data analysis application. Genome Biol 8:R6 2007.

13. Zhou G, Chen J, Lee S, Clark T, Rowley JD & Wang SM. The pattern of gene expression in human CD34$^+$ stem/progenitor cells. Proc Natl Acad Sci USA 98:13966-13971, 2001.

14. Georgantas RW, III, Tanadve V, Malehorn M, Heimfeld S, Chen C, Carr L, Martinez-Murillo F, Riggins G, Kowalski J & Civin CI. Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. Cancer Res 64:4434-4441, 2004.

15. Salvagiotto G, Zhao Y, Vodyanik M, Ruotti V, Stewart R, Marra M, Thomson J, Eaves C & Slukvin I. Molecular profiling reveals similarities and differences between primitive subsets of hematopoietic cells generated in vitro from human embryonic stem cells and in vivo during embryogenesis. Exp Hematol, 36(10):1377-89, 2008.

16. Raouf A, Zhao Y, To K, Stingl J, Delaney A, Iscove N, Jones S, Emerman J, Aparicio S, Marra M & Eaves C. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. Cell Stem Cell 3:109-118, 2008.

17. Score J, Curtis C, Waghorn K, Stalder M, Jotterand M, Grand FH & Cross NC. Identification of a novel imatinib responsive KIF5B-PDGFRA fusion gene following screening for PDGFRA overexpression in patients with hypereosinophilia. Leukemia 20:827-832, 2006.

18. Hogge DE, Lansdorp PM, Reid D, Gerhard B & Eaves CJ. Enhanced detection, maintenance and differentiation of primitive human hematopoietic cells in cultures containing murine fibroblasts engineered to produce human Steel factor, interleukin-3 and granulocyte colony-stimulating factor. Blood 88:3765-3773, 1996.

19. Nicolini FE, Holyoake TL, Cashman JD, Chu PPY, Lambie K & Eaves CJ. Unique differentiation programs of human fetal liver stem cells revealed both in vitro and in vivo in NOD/SCID mice. Blood 94:2686-2695, 1999.

20. Petzer AL, Eaves CJ, Lansdorp PM, Ponchio L, Barnett MJ & Eaves AC. Characterization of primitive subpopulations of normal and leukemic cells present in the blood of patients with newly diagnosed as well as established chronic myeloid leukemia. Blood 88:2162-2171, 1996.

SUPPORTING DATA:

Figure 1.

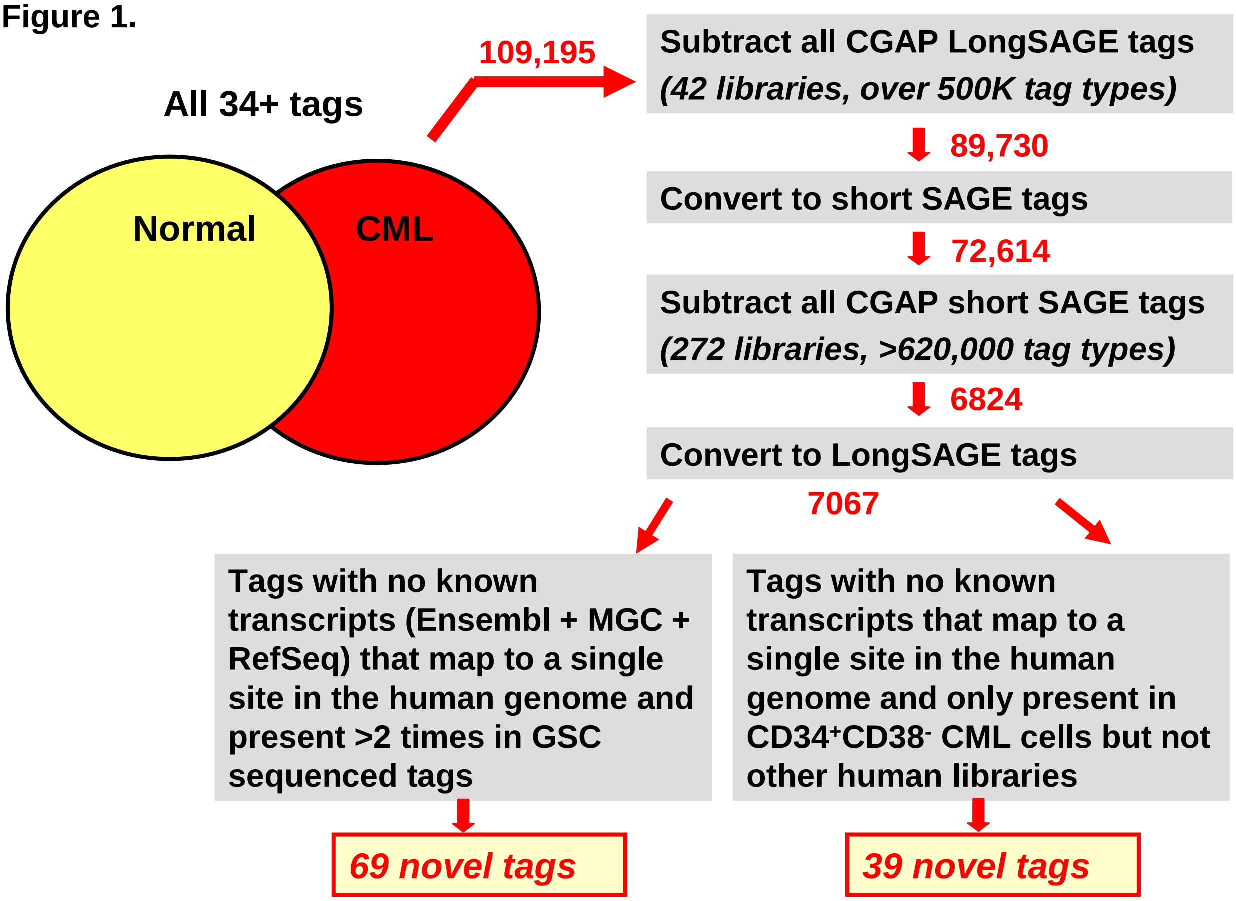

All 34+ tags
Normal
CML
109,195
Subtract all CGAP LongSAGE tags
(42 libraries, over 500K tag types)
89,730
Convert to short SAGE tags
72,614
Subtract all CGAP short SAGE tags
(272 libraries, >620,000 tag types)
6824
Convert to LongSAGE tags
7067
Tags with no known transcripts (Ensembl + MGC + RefSeq) that map to a single site in the human genome and present >2 times in GSC sequenced tags
Tags with no known transcripts that map to a single site in the human genome and only present in CD34+CD38- CML cells but not other human libraries
69 novel tags
39 novel tags

Figure 1. Identification of novel tags expressed uniquely in primitive CML cells. A CML meta-library made by pooling the lin$^-$CD34$^+$CD38$^+$ and lin$^-$CD34$^+$CD38- libraries generated from 3 individual CML patients was compared with a normal meta-library including lin$^-$CD34$^+$CD38$^+$ and lin$^-$CD34$^+$CD38$^-$ libraries generated from normal samples listed in Table 2.

Figure 2.

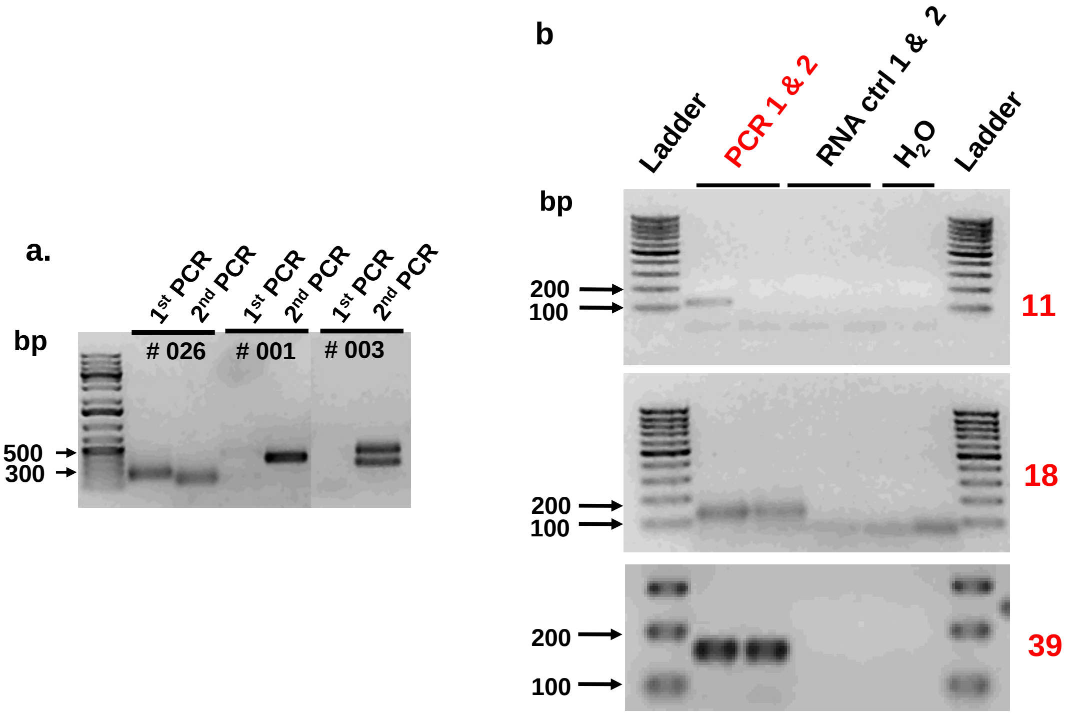


Figure 2. 3'-RACE (a) and PCR (b) results demonstrating detection of novel tags in cDNA preparations from lin$^-$CD34$^+$CD38$^-$ cells isolated from 3 different CML patients.

Figure 3

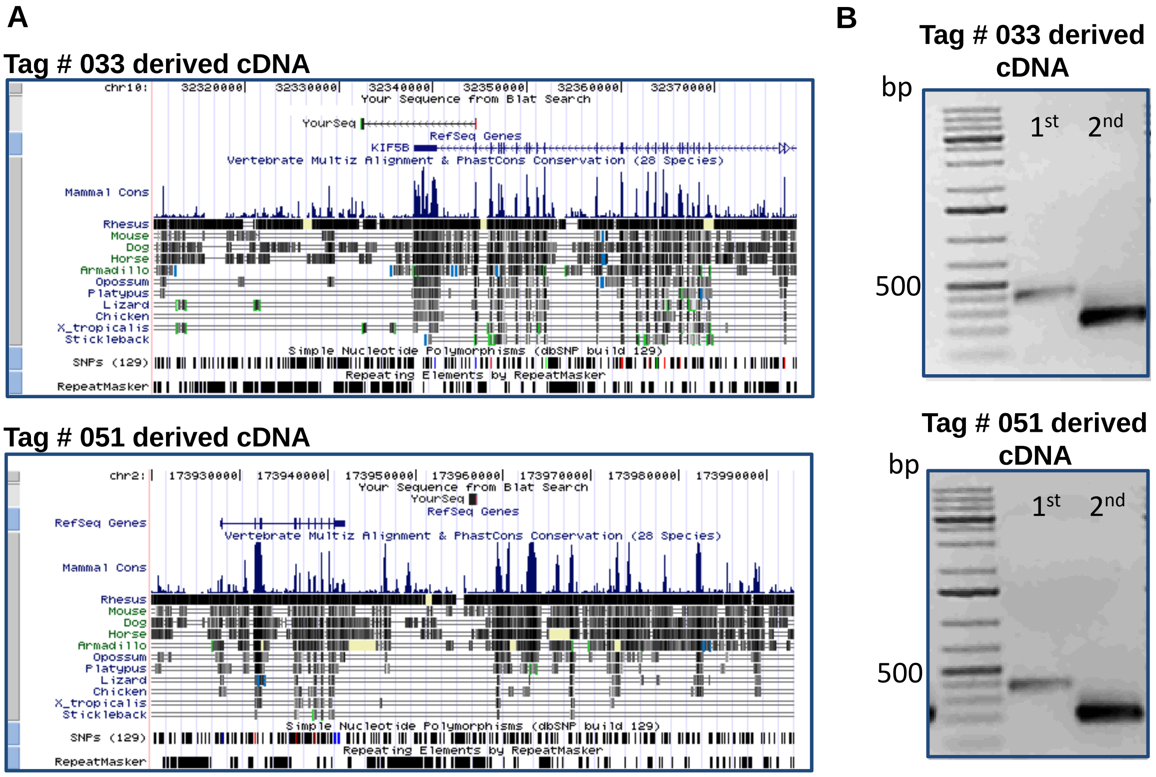

A
Tag # 033 derived cDNA
chr10: 32320000 32330000 32340000 32350000 32360000 32370000
Your Sequence from Blat Search
YourSeq
RefSeq Genes
KIF5B
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 129)
SNPs (129)
Repeating Elements by RepeatMasker
RepeatMasker
Tag # 051 derived cDNA
chr2: 173930000 173940000 173950000 173960000 173970000 173980000 173990000
Your Sequence from Blat Search
YourSeq
RefSeq Genes
RefSeq Genes
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 129)
SNPs (129)
Repeating Elements by RepeatMasker
RepeatMasker
B
Tag # 033 derived cDNA
bp
500
1st
2nd
Tag # 051 derived cDNA
bp
500
1st
2nd

Figure 3. (A) UCSC Genome Browser view of the novel tag derived cDNA fragments. Tag#033 derived cDNA provided an unknown 3'-UTR of KIF-5B; while Tag#051 derived cDNA located in an isolated genome region where the nearest RefSeq transcript is > 10kb away. (B) Nested PCR to confirm the validity of the tag derived cDNAs using primers designed with the sequencing data.

Figure 4

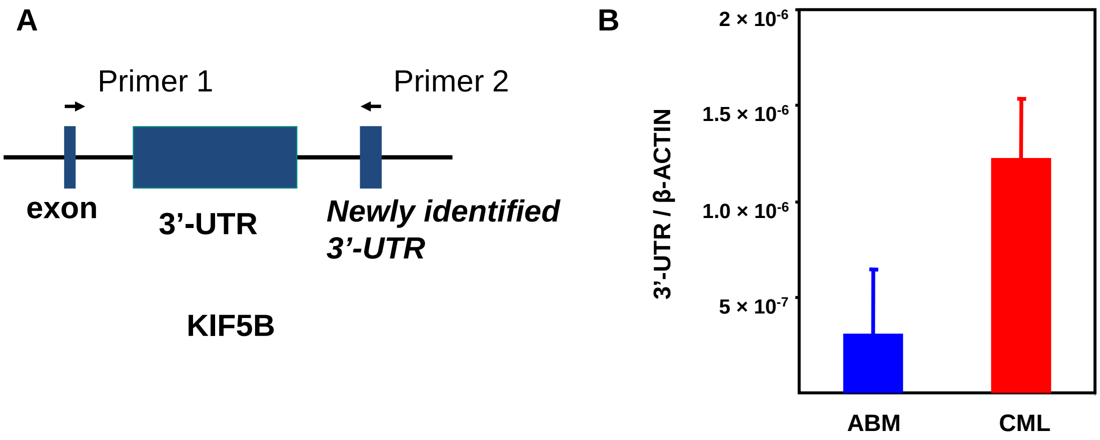


Figure 4. (A) Schematic structure of the newly identified 3’-UTR of KIF5B (B) The quantitative measurement of the newly identified 3’-UTR using lin-CD34+ cells of 2 ABM (adult bone marrow) vs. 9 samples from CML patients.

Table 1. Novel tags discovered in the transcriptome of normal adult human lin$^-$CD34$^+$ bone marrow cells

Tag #	Tag sequence	Tag counts	Chromosome localization
001**	TGACCAAATCCCCGTTT	2	3p22.1
002**	TTCCTAGATGGGAGGAC	2	10q11.22
003	AAAGTTACCTTCTATGT	1	8q21.3
004**	CTGCCACGCTACTTGTG	1	7p15.3
005	AAGTCGTCTTGTTTTGG	1	14q24.3
006	CCTGCTGGACCGTGGGG	1	6q15
007	CTACCAAAATGTAAAAG	1	11p13
008**	CTCTGTGACAGTCTAGA	1	Xq22.3
009**	GCACTAAAGAATCGTCA	1	9q21.2
010**	AGCGGAATAGAGAGAAG	1	10q22.3
011	CACAGCGACACTCTTGC	1	5p15.32
012	TTTAAGCTCTAATCTCT	1	1p21.1
013	ATAGCAAAGTCTAGAAC	1	10q21.1
014**	AACTTAGCCATTAGCTC	1	4q12
015	ACACCAAACACGATTAG	1	1p12
016**	ACACTTGCGGTAACAAA	1	17q25.3
017**	CATCAATTACCATCACT	1	9q22.32
018	CCTAGTTATCTACCCAA	1	8q24.12
019	TTATCCTTCTTCACCCC	1	15q25.2
020	TCCCGGGTGGTCCGGGT	1	11q23.3
021	GAGTCGTGTTTCTTATG	1	5p14.1
022**	TAAGTACCTACACAGTG	1	2q34
023	GTGTCATTAAATATGGG	1	4q25

Tag counts shown are absolute values from a total of 201,106 tags sequenced.
** These tags were validated by PCR in 3 RNA samples from normal adult human lin$^-$CD34$^+$ bone marrow cells.

Table 2. Summary of the LongSAGE libraries used

Cell Source	Phenotype	*In vitro* progenitor content		Total tags
		% CFC	% LTC-IC	
Adult bone marrow (Single donor)	[a] lin^-CD34^+	12	0.3	200,000
	[b] $lin^-CD34^+CD38^+$	25	1.5	200,000
	$lin^-CD34^+CD38^-$	14	39	200,000
Adult bone marrow (pool of 9)	$lin^-CD34^+CD38^+$	19	2	200,000
	$lin^-CD34^+CD38^-$	4	12	200,000
Mobilized peripheral blood (pool of 10)	$lin^-CD34^+CD38^+$	16	12	200,000
	$lin^-CD34^+CD38^-$	6	52	200,000
Umbilical cord blood (pool of ~150)	$lin^-CD34^+CD38^+$	15	4	200,000
	$lin^-CD34^+CD38^-$	6	14	200,000
Fetal liver (pool of 5)	$lin^-CD34^+CD38^-$	31	58	200,000
CML1	$lin^-CD34^+CD38^+$	17	1.3	200,000
	$lin^-CD34^+CD38^-$	8	16	200,000
CML2	$lin^-CD34^+CD38^+$	3	<0.01	200,000
	$lin^-CD34^+CD38^-$	4	0.2	200,000
CML3	$lin^-CD34^+CD38^+$	20	0.2	200,000
	$lin^-CD34^+CD38^-$	19	26	200,000

LTC-IC content was calculate by dividing the total CFCs detected after 6 wks of culture (per 100 cells initially seeded into the LTCs) by a predetermined average estimated CFC output per LTC-IC for each distinct source of cells (i.e., = 18, 25, 28[18], 72[19] and 8[20] for adult bone marrow, G-CSF mobilized peripheral blood, umbilical cord blood, fetal liver and CML cells, respectively).

[a] Lin refers to a cocktail of lineage (lin) markers, consisting of CD2, 3, 14, 16, 19, 20, 24, 56, 66b and Glycophorin A.

[b] Starting from this library, cells used to construct the LongSAGE libraries were also depleted of cells expressing CD7, 36, 45RA and 71.

Table 3. Tags unique to lin$^-$CD34$^+$ CML cell libraries

# of tag	Tags	CML1 38$^+$	CML1 38$^-$	CML2 38$^+$	CML2 38$^-$	CML3 38$^+$	CML3 38$^-$	Chromosome location
a 1	CAGCCTATTTACCAGAG	**0**	**0**	**0**	**0**	**0**	**6**	**3p13**
2	TAAGCTACATCCAGGAA	**0**	**0**	**0**	**0**	**2**	**3**	**1p36**
a 3	TAATGATGGGTACGGAG	**0**	**0**	**0**	**0**	**5**	**0**	**5q13**
4	TTGCCTTCGCGGAGGCC	**0**	**4**	**0**	**0**	**0**	**0**	**10p12**
a 5	CTTGGCCCTATCTCCAG	**0**	**0**	**0**	**0**	**3**	**0**	**15q26**
6	AGCAAAACGCTGTCTCA	**0**	**0**	**3**	**0**	**0**	**0**	**17p11**
7	AGTAGTTATTTAATAAT	**0**	**0**	**0**	**0**	**0**	**3**	**1p36**
8	CGTTTTGGGTCTTTTCC	**0**	**2**	**0**	**1**	**0**	**0**	**2p16**
9	TTATTACTTCTGTTGAT	**0**	**0**	**3**	**0**	**0**	**0**	**2q35**
10	AGACCCGGCAGGAGGAG	**0**	**0**	**0**	**0**	**3**	**0**	**3p24**
a 11	AGCGTGGCGCACAGCCC	**0**	**0**	**0**	**3**	**0**	**0**	**4q32**
12	GGCTCTGCATAAAAATT	**0**	**1**	**1**	**0**	**0**	**0**	**10p12**
13	CTCTTAGGATCAAATAT	**0**	**0**	**0**	**0**	**0**	**2**	**10q21**
14	GGTCGCAGGGAACTGTG	**0**	**1**	**0**	**0**	**1**	**0**	**11q13**
15	GACAGAGCACACATCAC	**0**	**0**	**0**	**0**	**0**	**2**	**11q22**
16	AATCCCTCTAGAATCTG	**0**	**0**	**2**	**0**	**0**	**0**	**12q13**
17	TCTTCGGCGCCTCTTCG	**0**	**0**	**0**	**0**	**2**	**0**	**14q11**
a 18	CAGTTACAGCATTTTCT	**1**	**1**	**0**	**0**	**0**	**0**	**14q32**
19	AAGCGTAACTGTGTGTG	**2**	**0**	**0**	**0**	**0**	**0**	**15q13**
20	CGGCATTTTTTTCGCTG	**0**	**0**	**2**	**0**	**0**	**0**	**15q26**
21	CCATTATCCCCTCCTGA	**2**	**0**	**0**	**0**	**0**	**0**	**19p13**
a 22	CACAAACCTCACAGACA	**0**	**0**	**0**	**0**	**0**	**2**	**20p13**
a,b 23	GTTCTCCGCCCTCCAGC	**0**	**0**	**0**	**0**	**1**	**1**	**21q22**
*** 24**	CTCAGTGCGGCCCTGGG	**0**	**2**	**0**	**0**	**0**	**0**	**2p21**
25	CAGGTCCCCGGTCGGAC	**0**	**0**	**2**	**0**	**0**	**0**	**4p16**
26	GATGGTAGACCACTTGG	**0**	**2**	**0**	**0**	**0**	**0**	**4q35**
27	ACAATTCTTAGACAGTA	**0**	**2**	**0**	**0**	**0**	**0**	**5q21**
28	ACTCCTTGACCGATGTA	**0**	**0**	**2**	**0**	**0**	**0**	**6p22**
29	CAATTTCGAATTACGAT	**1**	**1**	**0**	**0**	**0**	**0**	**6p24**
30	AACTCATCTAGATGCAT	**0**	**0**	**2**	**0**	**0**	**0**	**6q21**
31	GGCAACTCATCAAGATC	**0**	**2**	**0**	**0**	**0**	**0**	**7p13**
a,b 32	GTAGGATGGTGAAAATG	**0**	**0**	**0**	**0**	**0**	**2**	**9p24**
a,b 33	GCTTTCCAGTGCCCAGC	**1**	**0**	**0**	**0**	**0**	**0**	**10p11**
34	TAGGCAACATAGTGAGA	**0**	**1**	**0**	**0**	**0**	**0**	**11p14**
35	TGGAATAAGGAATGAAG	**1**	**0**	**0**	**0**	**0**	**0**	**11q12**
36	GCACCAGTACAGTAAGC	**0**	**0**	**1**	**0**	**0**	**0**	**11q13**
37	AAATTGGATGTTGTGCC	**0**	**0**	**0**	**0**	**0**	**1**	**12p11**
38	GTCCGCTGCCCAGTAAC	**1**	**0**	**0**	**0**	**0**	**0**	**12q24**
39	GATCTCCTTAAGGGTTC	**0**	**0**	**0**	**0**	**0**	**1**	**14q32**
40	TCACTCTGATGTGATGG	**1**	**0**	**0**	**0**	**0**	**0**	**16p11**
41	TGGCGCCACTGCATTCC	**1**	**0**	**0**	**0**	**0**	**0**	**16p12**
42	TCTCAGTGCAAACTCGA	**1**	**0**	**0**	**0**	**0**	**0**	**16q24**
43	TCCTGCGTTCCAGGCTT	**0**	**0**	**1**	**0**	**0**	**0**	**17q21**
44	CACAAGGCGTCTAGCTA	**0**	**1**	**0**	**0**	**0**	**0**	**18q11**
45	CGTGAGCGCTCGTGAAG	**0**	**0**	**0**	**1**	**0**	**0**	**18q23**
a,b 46	TGAGGACCTATGAGGAG	**0**	**0**	**0**	**1**	**0**	**0**	**19p13**
47	AACCGACAGATTCAGGA	**1**	**0**	**0**	**0**	**0**	**0**	**20p13**
48	ATCCTAGGATGTAGAAC	**0**	**1**	**0**	**0**	**0**	**0**	**20q13**

* 49	TCCTCAATGCGGCACTC	0	1	0	0	0	0	22q13.1
50	CCATCAATCTGTGTGTG	1	0	0	0	0	0	2p22
a,b 51	CTTATGCCAGATAGGAA	1	0	0	0	0	0	2q31
52	ATACTCAACTGCTTGAA	0	0	0	0	1	0	2q32
53	TTCTAAAGCATTTGTTC	0	0	0	1	0	0	2q33
54	TAGCTCTTTTCTTCCTC	0	0	1	0	0	0	3q27
55	TGTTGAACCCAGGGTTT	0	0	0	0	0	1	4p15
56	AGTAGATAAGGCTCTTT	0	0	0	0	1	0	4q21
57	AGAATCTATATGTATTA	0	0	0	1	0	0	5q14
58	AGCATAGTATTATGCTA	0	0	0	0	0	1	5q21
59	GCACTTGTCTTAGTTGT	0	1	0	0	0	0	5q34
60	GTTCCTATTGGATAATT	0	0	1	0	0	0	5q35
61	TTCTCGATGGACCTGGT	0	1	0	0	0	0	5q35
62	GGTGTATTGGTTTCCTA	0	0	1	0	0	0	6p22
63	ATCAGCCGGGTGTCGTG	0	1	0	0	0	0	6q12
64	TACTGCAAGACTCAGCA	0	0	0	0	0	1	6q25
65	AGTCTGACAGGGTTGCA	0	0	0	1	0	0	7p12
66	AACACCCCACCCCTTCC	1	0	0	0	0	0	9p13
67	TGAGAACCTATTAGGTC	1	0	0	0	0	0	9q22
68	CCAACCGCAACCTGGGA	0	0	0	0	1	0	9q34
69	GACTGTCTATTACTTGT	0	0	0	0	1	0	Xq26

Absolute tag counts for each library are shown.

[a.] Tag validated by the 3’-RACE and the derivative cDNA confirmed with sequencing

[b.] Tag validated by the 3’-RACE and the derivative cDNA confirmed with sequencing

* Tags present in Table 4.

Table 4. Tags unique to lin$^-$CD34$^+$CD38$^-$ CML cell libraries

# of tag	Tag	CML1	CML2	CML3	location
1	ACAATTCTTAGACAGTA	2	0	0	5q21.1
2	ACAGAACCATCCTGGGG	0	2	0	11p11.2
3	ACTTGAGTGAAACACTT	0	1	0	7p15.3
4	AGACAGTACAGAGCACA	0	1	0	12q24.13
5	ATCAGGCTTACTTTTTA	0	2	0	2p22.1
6	ATGATGTCTTCACATCA	0	1	0	Xp22.11
7	CAATACAGCTATTATTG	0	1	0	14q13.1
8	CACAAGGTTGGGCCCCC	0	1	0	2q14.3
9	CACGACGAAAGCCCTGG	0	1	0	14q32.31
10	CATAGTTTATGGACAGC	0	1	0	1p12
[a] 11	CATTCGTTCAACAAATG	0	1	0	11p11.2
12	CCAATTGGATAGACTTC	0	1	0	5q23.3
13	CCAGCTACGATCAGAGG	0	2	0	11q23.3
14	CCATTATTGGCAAGAAC	0	2	0	8q22.1
15	CCGGAAGGCTGGCCAGG	0	2	0	6p25.1
16	CGCTCATTACAGAACTG	0	2	0	8q13.1
17	CGTCCATCCTGGAAAGC	0	1	0	2p25.1
[a] 18	CTCAATGGCTGGAAGGC	0	1	0	8p22
*19	CTCAGTGCGGCCCTGGG	2	0	0	2q21.2
20	CTTTTGCCTAAAGCTCG	0	2	0	2q23.3
21	GACACAAACGCTGCTGC	0	1	0	11q23.2
22	GATAGGGTATATGGGTA	0	1	0	18q12.3
23	GATAGTGAGTATCAGTC	0	1	0	1q31.1
24	GATCTGGGGTTTCCCTA	0	1	0	12q24.21
25	GCGCCACTTCAGAGCCT	0	1	0	6p21.31
26	GGATCGCCAGCTTCTTT	0	1	0	10p11.22
27	GGGGTACATCCTCCTGC	0	2	0	13q31.1
28	GGTTACAGTTGTTTGTC	0	2	0	2p16.1
29	GGTTGTAAGCCCCACCT	0	2	0	2q22.2
30	GTAATGACATTGTGAAC	0	1	0	Xp11.4
31	GTCATTCCATAACCACC	0	1	0	15q22.31
32	GTTAGTATTAATGGAAG	0	1	0	9q21.13
33	TACCGTGGCTCACTTGG	0	2	0	8q12.2
34	TAGTAACTCTACTAGAT	0	1	0	13q31.1
35	TATTTGCTCTGAATTTT	0	1	0	5q13.3
*36	TCCTCAATGCGGCACTC	1	0	0	22q13.1
37	TCTGGAAGGGATTTTTG	0	1	0	3p22.3
38	TTCCCAGGCGGGGAGCG	0	2	0	7p22.3
[a] 39	TTTTCGAATCCCAACGC	2	0	0	3q22.3

Absolute tag counts for each library are shown.

[a.] Tags validated by a cDNA-dependent PCR approach

* Tags also present in Table 3.

APPENDICES:

Zhao Y, Raouf A, Kent D, Khattra J, Delaney A, Schnerch A, Asano J, McDonald H, Chan C, Jones S, Marra MA, Eaves CJ. A modified polymerase chain reaction-long serial analysis of gene expression protocol identifies novel transcripts in human CD34+ bone marrow cells. Stem Cells. 25:1681-9, 2007.

Zhao Y, Delaney A, Marra M, Eaves AC, Eaves CJ. Comparative transcriptome analysis of normal and chronic myeloid leukemia stem cells. Exp Hematol. 35 (Suppl. 2): p61, 2007.

Zhao Y, Delaney A, Marra M, Jiang X, Eaves AC, Eaves CJ. Comparative transcriptome analysis of different subsets of CD34+ normal and chronic myeloid leukemia cells identifies novel perturbations in the CML stem cell population. Blood. 110 (Suppl. 1): 19a, 2007.

Zhao Y, Delaney A, Raouf A, Raghuram K, Li HI, Schnerch A, Jiang X, Eaves AC, Marra MA, & Eaves CJ. Differentially expressed and novel transcripts in highly purified chronic phase CML stem cells. Blood. 112 (Suppl. 1): 79a, 2008.

STEM CELLS®

STEM CELL GENETICS AND GENOMICS

A Modified Polymerase Chain Reaction-Long Serial Analysis of Gene Expression Protocol Identifies Novel Transcripts in Human CD34$^+$ Bone Marrow Cells

YUN ZHAO,[a] AFSHIN RAOUF,[a] DAVID KENT,[a,b] JASWINDER KHATTRA,[c] ALLEN DELANEY,[c] ANGELIQUE SCHNERCH,[b,c] JENNIFER ASANO,[c] HELEN MCDONALD,[c] CHRISTINA CHAN,[a] STEVEN JONES,[c,d] MARCO A. MARRA,[c,d] CONNIE J. EAVES[a,d,e,f]

[a]Terry Fox Laboratory and [c]Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada; [b]Genetics Program and the Departments of [d]Medical Genetics, [e]Medicine, and [f]Experimental Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada

Key Words. Long serial analysis of gene expression • Stem/progenitor cells • Bone marrow • Transcriptome

ABSTRACT

Transcriptome profiling offers a powerful approach to investigating developmental processes. Long serial analysis of gene expression (LongSAGE) is particularly attractive for this purpose because of its inherent quantitative features and independence of both hybridization variables and prior knowledge of transcript identity. Here, we describe the validation and initial application of a modified protocol for amplifying cDNA preparations from <10 ng of RNA (<10^3 cells) to allow representative LongSAGE libraries to be constructed from rare stem cell-enriched populations. Quantitative real-time polymerase chain reaction (Q-RT-PCR) analyses and comparison of tag frequencies in replicate LongSAGE libraries produced from amplified and nonamplified cDNA preparations demonstrated preservation of the relative levels of different transcripts originally present at widely varying levels. This PCR-LongSAGE protocol was then used to obtain a 200,000-tag library from the CD34$^+$ subset of normal adult human bone marrow cells. Analysis of this library revealed many anticipated transcripts, as well as transcripts not previously known to be present in CD34$^+$ hematopoietic cells. The latter included numerous novel tags that mapped to unique and conserved sites in the human genome but not previously identified as transcribed elements in human cells. Q-RT-PCR was used to demonstrate that 10 of these novel tags were expressed in cDNA pools and present in extracts of other sources of normal human CD34$^+$ hematopoietic cells. These findings illustrate the power of LongSAGE to identify new transcripts in stem cell-enriched populations and indicate the potential of this approach to be extended to other sources of rare cells. STEM CELLS *2007;25:1681–1689*

Disclosure of potential conflicts of interest is found at the end of this article.

INTRODUCTION

Genome-wide expression profiling has become an important tool for analyzing cell behavior and has been particularly useful for identifying molecular events associated with early developmental decisions and disease pathogenesis. Two technologies are now commonly used for comparing or characterizing the complete transcriptome of specific cell populations: hybridization-based arrays [1] and serial analysis of gene expression (SAGE) [2]. In the first of these procedures, known transcript sequences or expressed sequence tags (ESTs) present on a solid phase surface were originally used to capture reverse-transcribed DNA copies of extracted cellular mRNA. The extent of hybridization achieved by competing cDNAs prepared from two cell sources was then determined to allow a comprehensive survey of differences in the gene expression profiles of the two cell populations being compared. The subsequent substitution of annotated oligonucleotides as capture probes has further improved consistency and signal detection.

SAGE involves the construction of large libraries of tags (typically 10 or 17 nucleotides long) that have been reverse-transcribed from the 3′ end of mRNAs present in the sample. The tags are then sequenced, and bioinformatics methods are used to derive transcript identities. Transcript levels can then be inferred directly from tag frequencies, bypassing any need for comparison to a reference cDNA preparation. As a result, each SAGE library becomes a permanent digital data resource accessible for repeated interrogation. The fact that SAGE does not require prior knowledge of the transcripts being surveyed also makes it useful for gene discovery. SAGE has thus become a particularly attractive technology for studies of cellular transcriptomes from organisms for which comprehensive genomic sequence information is available. Nevertheless, a major limitation of the original SAGE methodology has been the need for relatively large quantities of starting RNA (originally 5 μg, the amount typically obtained from approximately 10^6 cells [2]).

Correspondence: Connie. J. Eaves, Ph.D., Terry Fox Laboratory, 675 West 10th Avenue, Vancouver, BC, Canada V5Z 1L3. Telephone: 604-675-8122; Fax: 604-877-0712; e-mail: ceaves@bccrc.ca Received December 6, 2006; accepted for publication March 26, 2007; first published online in STEM CELLS *EXPRESS* April 5, 2007; available online without subscription through the open access option. ©AlphaMed Press 1066-5099/2007/$30.00/0 doi: 10.1634/stemcells.2006-0794

Subsequent modifications to decrease the amount of starting material needed (microSAGE [3], amplified antisense RNA-LongSAGE [4], small amplified RNA-SAGE [5], SAGE-lite [6], and polymerase chain reaction [PCR]-SAGE [7]) have now made it possible for either SAGE or LongSAGE libraries to be generated from much smaller amounts of RNA (down to 40 ng). However, these are still not readily applicable to isolates containing fewer than 10^4 cells. Because of the very low frequency of normal or malignant stem cells in many primary tissues, this limitation still hampers the use of any SAGE approach for characterizing a variety of stem cell populations.

Here, we describe a method that adapts recent technology for amplifying cDNAs from a few nanograms of total cellular RNA [8, 9] in a fashion that meets the requirements for SAGE library construction, minimizes the generation of ambiguous tags, and preserves the initial transcript representation. Using this approach, we have created the first LongSAGE library thus far reported from the $CD34^+$ subset of normal adult human bone marrow cells. Analysis of the tags obtained indicates the capture of many expected transcripts, as well as a number of transcripts not previously known to exist.

Materials and Methods

Cells

Normal human cord blood cells were obtained, with consent, from anonymized discarded placentas, and the low-density (<1.077 g/cm^3) fraction of cells isolated by centrifugation on Ficoll-Hypaque (Pharmacia, Calgary, AB, Canada, http://www.pfizer.ca) was then cryopreserved. Samples were thawed, and the $CD34^+$ cells were separated immunomagnetically using a $CD34^+$ cell positive selection kit (EasySep; Stem Cell Technologies, Vancouver, BC, Canada, http://www.stemcell.com). The cells were then stained with a phycoerythrin-conjugated anti-human CD34 antibody (8G12; BD Biosciences [BD], San Jose, CA, http://www.bdbiosciences.com) and propidium iodide (PI) (Sigma-Aldrich, St. Louis, http://www.sigmaaldrich.com), and a population of viable (PI^-) $CD34^+$ cells was obtained using a FACSVantage machine (BD). Aliquots of 100, 500, 10^3, and 10^5 viable $CD34^+$ cells were collected directly into vials containing 100 μl of RNA extraction buffer from the PicoPure RNA extraction kit (Arcturus, Mountain View, CA, http://www.arctur.com). Cryopreserved normal adult human bone marrow cells obtained with informed consent were provided by the Northwest Tissue Center (Seattle). After thawing, human cells expressing lineage (lin) markers of mature blood cells (CD2, CD3, CD14, CD16, CD19, CD24, CD56, CD66b, and glycophorin A) were removed immunomagnetically using a column (StemSep; Stem Cell Technologies) as recommended by the manufacturer and cryopreserved. The lin^- cells were thawed at 37°C and incubated in 50% fetal calf serum in Hanks' balanced salt solution overnight at 4°C to minimize effects of freezing and thawing on the levels of different mRNAs present. This was established by comparing the levels of transcripts for 11 variably expressed genes using quantitative real-time (Q-RT)-PCR. We also found that the gentle thawing process adopted for previously cryopreserved cells did not perturb the differentiation capabilities of these cells as determined by colony-forming cell (CFC) assays (data not shown). Thawed cells were then stained with allophycoerythrin-conjugated anti-human CD34 antibody (8G12; BD), fluorescein isothiocyanate-conjugated lineage marker antibodies, and PI. Viable (PI^-) lin^-CD34^+ cells were then isolated using a FACSVantage flow cytometer. The purity of the sorted cells was determined to be $>98\%$ as assessed by second fluorescence-activated cell sorting (FACS) analysis of an aliquot of the sorted cells. Total RNA extracts were prepared from viable lin^-CD34^+ cells isolated by FACS in the same manner as described for cord blood cells.

Hematopoietic Progenitor Cell Assays

CFC assays were performed by plating human lin^-CD34^+ bone marrow cells at 800 cells per milliliter in serum-containing methylcellulose medium (Methocult 4230; Stem Cell Technologies) supplemented with 3 U/ml erythropoietin (Stem Cell Technologies), 50 ng/ml Steel factor (SF) (prepared and purified in the Terry Fox Laboratory), 20 ng/ml each of interleukin-3 (IL-3) and granulocyte-macrophage colony-stimulating factor (both from Novartis International, Basel, Switzerland, http://www.novartis.com), 20 ng/ml granulocyte colony-stimulating factor (G-CSF) (Stem Cell Technologies), and 20 ng/ml IL-6 (Cangene Corp., Mississauga, ON, Canada, http://www.cangene.com) [10]. Long-term culture-initiating cell (LTC-IC) assays were performed by culturing 2×10^4 lin^-CD34^+ bone marrow cells in 2 ml of myeloid LTC medium (Myelocult; Stem Cell Technologies) supplemented with 10^{-6} mol/l hydrocortisone sodium hemisuccinate (Sigma-Aldrich) for 6 weeks on pre-established, irradiated feeder layers of mouse fibroblasts genetically engineered to produce human SF, G-CSF, and IL-3. At the end of this time, the number of CFCs present was determined, and the number of input LTC-IC calculated assuming an average 6-week output of 18 CFCs per LTC-IC [10].

RNA Isolation and cDNA Preparation and Amplification

An RNA extract prepared from undifferentiated H9 human embryonic stem cells was kindly provided by Dr. J. Thomson (University of Wisconsin, Madison, WI). RNA extracts were also prepared separately from 100, 500, 10^3, or 10^5 FACS-purified human $CD34^+$ cord blood cells using the PicoPure RNA extraction kit. To minimize contamination with genomic DNA, RNA isolates were treated with DNaseI (Amplification Grade; Invitrogen, Burlington, ON, Canada, http://www.invitrogen.com) according to the manufacturer's protocol. To quantify the extent of genomic contamination in the final purified cDNA used for SAGE library construction, we used intron-specific primers to amplify sequences for two genes on different chromosomes: forward primer 5′-CCCCATGAGTCAGGTCGG-3′ and reverse primer 5′-CCCAGACTGCATCTCAGCCA-3′ for the *DRCG8* gene (22q11.2), and forward primer 5′-AGTTTCTCCTCTCTCCTCCCAAG-3′ and reverse primer 5′-TCACTTCACTTCATTTTCACTTCTC-3′ for the *ATP11A* gene (13q34), by quantitative PCR (Q-PCR). The results obtained with both pairs of primers showed that $<0.1\%$ of the cDNA sample contained genomic DNA. RNAs were reverse-transcribed, and the cDNAs obtained were amplified using the switching mechanism at the 5′ end of RNA transcripts (SMART) cDNA synthesis kit (catalog number 635000; Clontech, Mountain View, CA, http://www.clontech.com) following the manufacturer's protocol but using modified template switching (TS) and cDNA amplification primers, as detailed below (Fig. 1A). The first-strand cDNA was synthesized with an oligo(dT) primer (5′-AAG CAG TGG TAA CAA CGC AGG CTA CTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TVN-3′, where V denotes A, C, or G and N denotes A, C, G, or T) and the PowerScript reverse transcriptase provided in the kit, in the presence of the modified TS primer. The TS primer was modified by introducing a sequence containing an *Asc*I digestion site (5′-AAG CAG TGG TAA CAA CGC A<u>GG CGC GCC</u> GGG-3′ [the *Asc*I site is underlined]). The first-strand cDNA was purified with a NucleoSpin column and then amplified using a modified PCR primer that contained a biotin molecule at its 5′ end (5′-biotin-AAG CAG TGG TAA CAA CGC AGG C-3′) and the Advantage II PCR Kit (Clontech). The biotinylated 5′ ends of the amplified cDNAs were then removed by digestion of the initial amplified product with *Asc*I (New England Biolabs, Beverly, MA, http://www.neb.com). The cDNA was purified on a Chroma-Spin 200 Column (Clontech), and its concentration was determined using a spectrophotometer (GeneQuant Pro; Biochrom, Cambridge, U.K., http://www.biochrom.co.uk).

SAGE Library Construction

For the PCR-LongSAGE libraries, the amplified cDNA was first digested with *Nla*III and incubated with streptavidin beads (M-280; Invitrogen); the immobilized, truncated cDNAs were then linked to

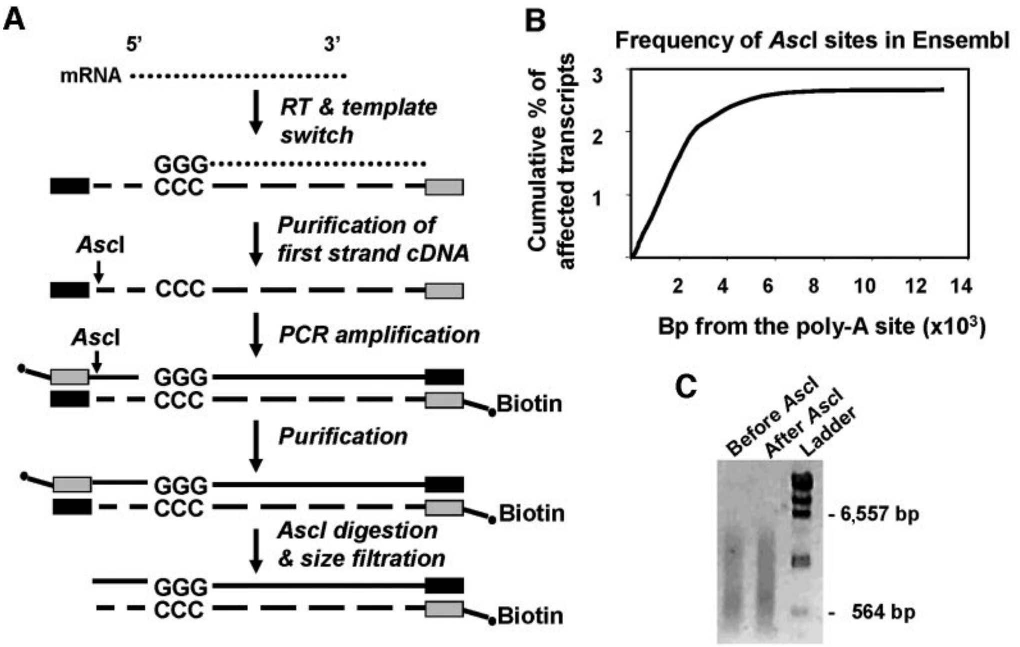


Figure 1. cDNA amplification protocol for PCR-serial analysis of gene expression (PCR-SAGE) library generation. **(A):** Incorporation of a template switching primer containing an *Asc*I sequence allows an end-to-end amplification of the first-strand cDNA using a single biotinylated oligonucleotide primer and then subsequent removal of the 3′ biotin via *Asc*I digestion. The 5′ end of the double-stranded cDNA is then available for capture on streptavidin-coated beads for SAGE library construction. **(B):** Rationale for choosing *Asc*I to eliminate the 3′-biotinylated end of amplified cDNAs based on the low frequency of its recognition sequences in human cDNAs. **(C):** The amplified cDNA smear before and after *Asc*I digestion. Amplified cDNA prepared from 10 ng of RNA was subjected to digestion with *Asc*I restriction endonuclease for 1 hour and size-fractionated on an ethidium bromide-stained agarose gel in parallel with undigested amplified cDNA sample. The results demonstrate that digestion with *Asc*I does not perturb the overall distribution of the amplified cDNA fragment size. Abbreviations: bp, base pairs; PCR, polymerase chain reaction; RT, reverse transcription.

two different adaptors, and LongSAGE libraries were constructed using the I-SAGE kit (Invitrogen) following the manufacturer's protocol. The I-SAGE kit was also used to construct a SAGE-lite library from 400 ng of PCR-amplified cDNA (22 cycles) obtained using a methodology described before [6] that also uses SMART cDNA technology.

Q-RT-PCR

RNA was reverse-transcribed with SuperScriptII (Invitrogen) to generate first-strand cDNA for use as the template for Q-RT-PCR analysis of transcript levels in nonamplified RNA preparations. Q-RT-PCR was performed using SYBR Green PCR MasterMix (Applied Biosystems, Foster City, CA, http://www.appliedbiosystems.com) and an iCycler PCR machine (Bio-Rad, Hercules, CA, http://www.bio-rad.com). After an initial denaturation step at 94°C for 5 minutes, 50 cycles of a three-step PCR with a single fluorescence measurement were undertaken (94°C for 15 seconds, 60°C for 20 seconds, and 72°C for 30 seconds). The PCR products were also subjected to melting curve analysis for verification of single amplicons and absence of primer dimers. Q-RT-PCR and data analysis were performed on an iCycler iQ system, using iCycler iQ real-time detection software (Bio-Rad). The primers used are shown in supplemental online Table 1. Q-PCR assays were used to confirm the expression of the unique tags identified by bioinformatics analysis of the lin$^-$CD34$^+$ human adult bone marrow LongSAGE library. For this purpose, RNA was extracted from lin$^-$CD34$^+$ cells isolated from human bone marrow samples from three different normal adult donors; one of these samples was the same as that used for construction of the PCR-LongSAGE library. cDNA was generated, as described and as a negative control the same amount of RNA was used without adding reverse transcriptase. Primers for detecting novel transcripts were selected from the human genome (Human BLAT Search, http://genome.ucsc.edu/cgi-bin/hgBlat) flanking 5′ and 3′ regions of the identified unique tags in such a way that the amplicons would include the unique tag sequences (supplemental online Table 2).

Bioinformatics and Statistical Methods

DiscoverySpace software (http://www.bcgsc.ca/platform/bioinfo/software/ds) was used to determine the similarity of different pairs of LongSAGE libraries using Audic-Claverie statistics [11] and for tag-to-gene mapping using the RefSeq database (build 35, August 26, 2004; http://www.ncbi.nlm.nih.gov/RefSeq). Pearson correlation coefficients were calculated using the regression program from the !STAT package [12], and hierarchical clustering was performed using Phylip software [13; http://www.med.nyu/rcr/phylip.main.html].

RESULTS

Development of a cDNA Amplification Protocol Suitable for Constructing LongSAGE Libraries

To allow LongSAGE libraries to be constructed from highly PCR-amplified preparations of 3′ cDNAs without major distortion of the original transcript representation, we used the SMART technology developed by Clontech [9] and also used in the SAGE-lite protocol [6] with two modifications. The original technology makes use of a TS primer containing a short polyguanine sequence at its 3′ end for the first-strand cDNA synthesis step. We then modified the cDNA amplification primer so that it contained a biotin molecule at the 5′ end. In addition, we modified the TS primer by introducing an eight-base (GGCGCGCC) *Asc*I restriction endonuclease recognition sequence into its 3′ end (Fig. 1A). These modifications allowed the biotinylated primers incorporated into the 3′ ends of the cDNA products to later be removed to yield a final product in which the cDNAs were biotinylated exclusively at their 5′ ends, as required for SAGE library construction (Fig. 1A). This approach is a variation of the previously described introduction of

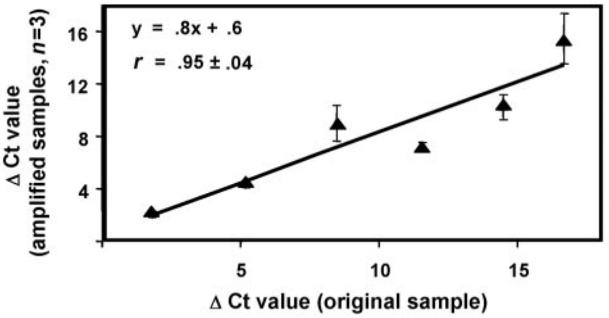


Figure 2. Real-time polymerase chain reaction (PCR) of replicate amplified cDNA samples. Three 10-ng aliquots of RNA extracted from a single pool of undifferentiated H9 human ESCs were independently amplified using a 21-cycle PCR step. Levels of seven transcripts (*ACTB, GAPD, BLP1, SAFB2, CCND1, ABCG2,* and *RPS4X*) were quantified by real-time PCR, and the values were normalized to the levels of *ACTB* transcripts measured in the same preparations. Q-RT-PCR data were also obtained from an initial aliquot of 100 ng of the same RNA after reverse transcription but with no amplification. Values shown are the mean ± SEM. The sequences of the primers used are shown in supplemental online Table 1. Abbreviation: Ct, cycle threshold.

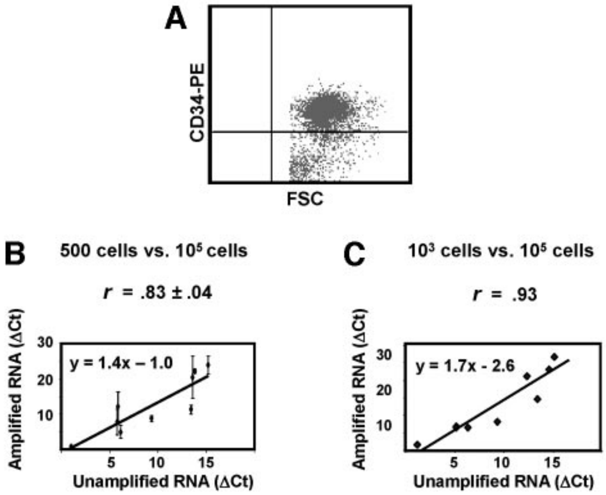


Figure 3. Validation of the applicability of the cDNA amplification procedure to small-cell numbers. **(A):** Fluorescence-activated cell sorting (FACS) profile showing the immunomagnetically enriched CD34$^+$ low-density human cord blood cells from which the final CD34$^+$ cells used in this study were isolated by FACS. **(B):** cDNA products were generated individually from three replicate aliquots of 500 CD34$^+$ cells sorted directly into RNA lysis buffer and then subjected individually to our modified cDNA amplification procedure. The levels of 11 transcripts (*GAPD, ACTB, ABCG2, CCND2, ABCB1, CCND1, BCR, ABL1, PIAS4, CD34,* and *SELL*) in each of the amplified cDNA samples were then quantified by quantitative real-time polymerase chain reaction and normalized to the levels of *ACTB* cDNA in the same samples. (The sequences of the primers used are shown in supplemental online Table 1). Values shown are the mean ± SEM. Pearson correlation coefficients and the best line fit to the data derived by least squares analysis are shown. **(C):** Similar analysis of RNA from replicate samples of 10^3 CD34$^+$ cells. Abbreviations: Ct, cycle threshold; FSC, forward light scattering; PE, phycoerythrin.

a seven-base *Sap*I site for the same purpose [7]. However, from in silico analyses, we found that 24% of Ensembl transcripts contain at least one *Sap*I site, which could result in a potential loss of >600 tag types following *Sap*I digestion. In contrast, only 3% of Ensembl transcripts contain one or more *Asc*I restriction sites, and only 80 contain an *Asc*I site between the first *Nla*III site 5′ of the poly(A) tail and the poly(A) tail itself (Fig. 1B). Consistent with the expectation of a minimal loss of tags after digestion with *Asc*I (at 37°C for 1 hour), we found that there was no detectable change in the size distribution of the amplified cDNAs when they were analyzed electrophoretically (Fig. 1C).

To determine the number of cycles of amplification to use, we generated cDNA samples independently from three separate 10-ng aliquots of RNA extracted from undifferentiated human H9 embryonic stem cells (http://www.transcriptomES.org) and then examined the electrophoretically separated products obtained after 18–24 cycles of amplification. The results showed that the PCR amplification reaction had not yet reached a plateau after 21 cycles, by which time there was already sufficient product to construct a SAGE library (supplemental online Fig. 1A). This result was also validated by Q-RT-PCR analyses (supplemental online Fig. 1B).

Evidence of the reproducibility of the cDNA amplification protocol and its ability to preserve relative transcript levels in amplified cDNA products was obtained from separate Q-RT-PCR measurements of the levels of six differentially expressed mRNAs in the H9 cell extract described above on samples taken before and after three independent amplifications of the starting cDNA pool (Fig. 2).

We next asked what would be the minimum number of normal adult human cells from which a suitable amplified cDNA product could be obtained to allow construction of a 200,000-tag LongSAGE library. To address this question, we used a combination of immunomagnetic cell separation and multiparameter FACS procedures to isolate CD34$^+$ cells from a pool of cells from three normal human cord blood harvests (Fig. 3A). cDNA products were prepared from separately collected aliquots of 100, 500, 10^3, and 10^5 of the CD34$^+$ cells isolated, and they were then amplified or not (10^5 cell samples). Figures 3B and 3C show comparisons of the levels of 10 transcripts quantified in these extracts by Q-RT-PCR before and after amplification. All 10 transcript species were detected in the amplified cDNA products obtained from as few as 500 cells, and their levels were highly correlated with those measured in the nonamplified material ($R = 0.83 \pm 0.04$; Fig. 3B). In addition, the RNA extracted from the 500-cell sample yielded more than 400 ng of amplified cDNA, which is more than enough to build a one million-tag LongSAGE library using the I-SAGE protocol (Invitrogen). The cDNA products generated from 100 CD34$^+$ cord blood cells also showed a significant correlation between the levels of the more prevalent transcript species before and after their amplification, although some of the rarer transcript species were not detectable in the amplified products generated in this case (data not shown).

Comparison of Replicate LongSAGE Libraries Prepared from Amplified and Nonamplified cDNAs

We then compared the complete tag profiles from LongSAGE libraries constructed from amplified and nonamplified cDNAs derived from the same original RNA extract. For this analysis, two of the independently amplified H9 cDNA preparations analyzed in Figure 2 were used to prepare replicate libraries. The two PCR-LongSAGE libraries were sequenced to depths of 57,470 (library A) and 112,517 (library B) total tags (all analyses performed using http://www.transcriptomES.org). To minimize effects due to poor-quality tags, we applied sequence quality cut-offs of 95.0% and 99.9% to the nonsingleton and singleton tags, respectively. This reduced the number of tags in the two PCR-LongSAGE libraries to 46,241 (library A) and 83,557 (library B). The library prepared from nonamplified material was a 467,522-tag library constructed from 20 μg of RNA using the standard I-SAGE protocol. Also included in this

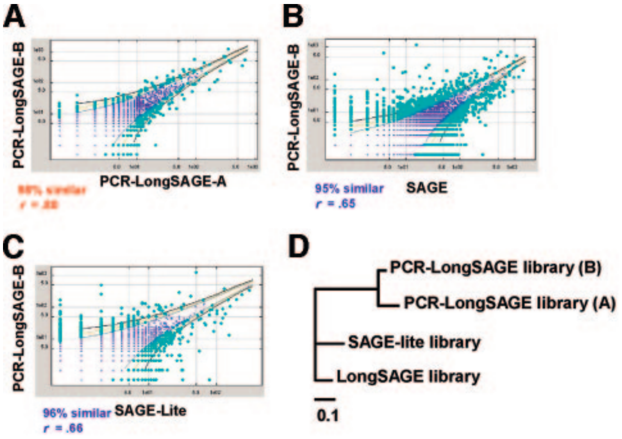
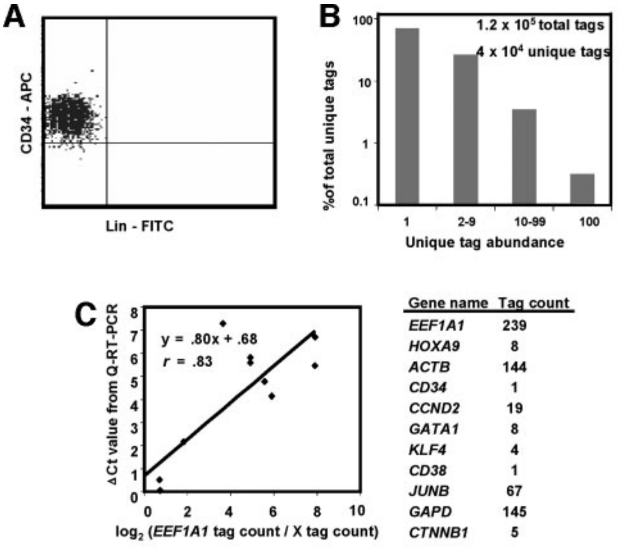


Figure 4. Comparisons of four LongSAGE libraries prepared from the same original RNA sample using different protocols. **(A):** Comparison using DiscoverySpace software of PCR-LongSAGE-B and PCR-LongSAGE-A. **(B):** Comparison using DiscoverySpace software of PCR-LongSAGE-B and the LongSAGE library constructed from the same RNA without prior amplification. **(C):** Comparison using DiscoverySpace software of PCR-LongSAGE-B and the SAGE-lite library constructed from the same RNA. **(D):** Hierarchical cluster analysis of the same four LongSAGE libraries demonstrates the similarity of the two PCR-LongSAGE libraries indicative of the reproducibility of the amplification process. Abbreviations: LongSAGE, long serial analysis of gene expression; PCR, polymerase chain reaction; SAGE, serial analysis of gene expression.

analysis was a 60,492-tag SAGE-lite library prepared from a 100-ng aliquot of the same RNA extract. All four libraries showed the expected predominance of low-abundance tags and, in this respect, were indistinguishable from one another (data not shown). They also contained readily detectable frequencies of tags unique to transcripts of known relevance to undifferentiated human embryonic stem cells (supplemental online Table 3) [14].

We then used DiscoverySpace software to compare the tag representation in these four libraries on a pairwise basis. This software uses Audic-Claverie statistics [11] to allow the tag composition of SAGE libraries to be compared independent of library size. This analysis showed the two replicate PCR-LongSAGE libraries to be 98% similar to one another using a 95% confidence interval, that is, only 2% of tag types were present at significantly different levels ($p < .05$) in one of the two PCR-LongSAGE libraries (Fig. 4A). Comparison of each of these libraries to the conventional LongSAGE library prepared from nonamplified material gave corresponding similarity values of 95% (for PCR-LongSAGE library B; Fig. 4B) and 84% for the PCR-LongSAGE library A (data not shown). Values for parallel similarity comparisons with the SAGE-lite library were 96% (library B; Fig. 4C) and 97% (library A; data not shown), and the value for comparison of the LongSAGE library with the SAGE-lite library was 97% (data not shown). In fact, only seven tags were consistently over- or under-represented in both of the PCR-LongSAGE libraries compared with the tags from the LongSAGE library prepared from nonamplified material, and none of these mapped to a unique site in the most recent version of the human genome (RefSeq database, build 35, August 26, 2004).

Pearson correlation analysis of tag frequencies in each pair of libraries generated correlation coefficients of 0.8 for the two PCR-LongSAGE libraries and somewhat lower values when these were compared with the library obtained from nonamplified material (0.61 and 0.65, respectively) or to a corresponding SAGE-lite library (0.61 and 0.66, respectively) (Fig. 4D). This latter method of comparison is more sensitive to differences between higher frequency tags. Hence, to avoid distortion from repetitive sequences, only tags that could be matched to a unique sequence in the most recent version of the human genome (build 35, August 26, 2004) were included in this analysis.



Gene name	Tag count
EEF1A1	239
HOXA9	8
ACTB	144
CD34	1
CCND2	19
GATA1	8
KLF4	4
CD38	1
JUNB	67
GAPD	145
CTNNB1	5

Figure 5. Description and validation of PCR-long serial analysis of gene expression (PCR-LongSAGE) library from human $\text{lin}^-\text{CD34}^+$ adult bone marrow cells. **(A):** Fluorescence-activated cell sorting (FACS) plot demonstrating the high purity (98%) of the $\text{lin}^-\text{CD34}^+$ cells isolated by two successive re-sorts of lin^- low-density normal adult human bone marrow cells and reanalyzed in a third run of the cells through the FACS instrument. **(B):** Distribution of tags in the PCR-LongSAGE library generated from the RNA extracted from these cells. **(C):** Ten transcripts identified from the PCR-LongSAGE library were chosen to test the correlation between PCR-LongSAGE and Q-RT-PCR methodologies. $\text{lin}^-\text{CD34}^+$ cells were isolated from the same donor, and Q-RT-PCR was performed on the cDNA products obtained. EEF1A1 was the most abundantly expressed transcript of those analyzed (based on SAGE tag counts) and was therefore chosen as a standard against which the other nine transcripts were compared. The *y*-axis shows the ΔCt value obtained in each case from the Q-RT-PCR measurements ($\Delta\text{Ct} = \text{Ct}_{(X)} - \text{Ct}_{(EEF1A1)}$), and the *x*-axis shows the corresponding tag frequency expressed as a $\log_2$ value after normalization against the EEF1A1 tag frequency ($\log_2$ [EEF1A1 tag count/X tag count]). Abbreviations: APC, allophycoerythrin; Ct, cycle threshold; FITC, fluorescein isothiocyanate; Q-RT-PCR, quantitative real-time polymerase chain reaction.

Construction and Analysis of a PCR-LongSAGE Library from CD34^+ Cells Isolated from Normal Adult Human Bone Marrow

We then used this method to construct a library from ~3,000 highly purified $\text{lin}^-\text{CD34}^+$ cells isolated by FACS from a sample of normal adult human bone marrow cells (Fig. 5A). Functional assays applied to these CD34^+ cells demonstrated that 12% had granulopoietic, erythroid, or mixed granulopoietic and erythroid CFC activity in vitro. In addition, 0.3% of these cells were detectable as 6-week precursors of CFCs in LTC-IC assays [10], as described in Materials and Methods. From this library, 201,106 tags were sequenced, and 42,310 unique tag types were obtained with a typical SAGE tag frequency distribution (Fig. 5B). A complete listing of all the tags is given at http://www.transcriptomES.org. Q-RT-PCR of cDNA preparations generated from extracts of independently purified $\text{lin}^-\text{CD34}^+$ cells from the same bone marrow sample showed a good correlation between the transcript levels measured and

Table 1. Transcripts detected in a polymerase chain reaction-long serial analysis of gene expression library prepared from normal adult human lin$^-$CD34$^+$ bone marrow cells

Symbol	Accession no.	Gene name	Location	Counts	Tag	Position
Transcription factor						
RUNX1	NM_001754	Runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)	21q22.3	2	AGAGAATATCCCAGAAC	1
				2	GTTCCAATTTTTTTTAA	2
LMO2	NM_005574	LIM domain only 2 (rhombotin-like 1)	11p13	13	GAGACGCATTTCGGTTG	1
ETV6	NM_001987	ets variant gene 6 (TEL oncogene)	12p13	2	CCAAGTGAACATTCTTG	1
HLF	NM_002126	Hepatic leukemia factor	17q22	1	GACCATCCAAATTTATG	1
PCGF4	NM_005180	B lymphoma Mo-MLV insertion region (mouse)	10p11.23	2	TTTGTATGGGAAAATTG	1
GATA2	NM_032638	GATA binding protein 2	3q21.3	4	GACAGTTGTTTGGAGAA	1
				1	GGCTAGGGAACAGATGG	2
GATA1	NM_002049	GATA binding protein 1 (globin transcription factor 1)	Xp11.23	10	GCCTCCAGAGGAGGGGT	1
MYB	NM_005375	v-myb myeloblastosis viral oncogene homolog (avian)	6q22–q23	12	GATCCTGTGTTTGCAAC	1
FLI1	NM_002017	Friend leukemia virus integration 1	11q24.1–q24.3	3	TTGTAAAATAATTTGAC	1
				1	TTCTGGTTTGAGATTTA	2
XBP1	NM_005080	X-box binding protein 1	22q12.1	14	CAATTAAAAGGTACAAT	1
CEBPA	NM_004364	CCAAT/enhancer binding protein (C/EBP), α	19q13.1	1	GGGGGTGAAGGGCCACT	1
Cell membrane						
CD34	NM_001773	CD34 antigen	1q32	1	GCTTCCTCCTCCCTCCT	1
FLT3	NM_004119	fms-related tyrosine kinase 3	13q12	1	GGAATTCATTTCACTCT	2
CSF3R	NM_172313	Colony stimulating factor 3 receptor (granulocyte)	1p35–p34.3	12	CTCCATCCAGCCCCACC	1
KIT	NM_000222	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	4q11–q12	3	AGTCCTTGAAAATATTT	1
EPOR	NM_000121	Erythropoietin receptor	19p13.3–p13.2	1	GACACTGTGCCCTGAGC	1
NOTCH1	NM_017617	Notch homolog 1, translocation-associated (*Drosophila*)	9q34.3	3	AGGAACTGTAGATGATG	1
CXCR4	NM_001008540	Chemokine (C-X-C motif) receptor 4	2q21	14	TTAAACTTAAAAAAAAA	1
CD44	NM_000610	CD44 antigen (homing function and Indian blood group system)	11p13	9	ATATGTATATTGCTGAG	1
SELL	NM_000655	Selectin L (lymphocyte adhesion molecule 1)	1q23–q25	9	CAATTTTGCATTTGAAT	1
PROM1	NM_006017	Prominin 1	4p15.32	5	TGCAGATTGCAGTTCTG	1
Kinase activity and signal adaptor						
JAK2	NM_004972	Janus kinase 2 (a protein tyrosine kinase)	9p24	2	TACTGTAAATATTTTTC	2
LYN	NM_002350	v-yes-1 Yamaguchi sarcoma viral-related oncogene homolog	8q13	4	ACATTTCTTTGTGCTTT	1
LNK	NM_005475	Lymphocyte adaptor protein	12q24	2	CTTTCTATTCAGGACTA	1
ALDH						
ALDH1B1	NM_000692	ALDH 1 family, member B1	9p11.1	1	AATTAACTCCGTTAAAA	1
ALDH3A2	NM_000382	ALDH 3 family, member A2	17p11.2	1	AGCCTGTGTTCCCAGCT	3
ALDH4A1	NM_170726	ALDH 4 family, member A1	1p36	1	AGGGGCCGGGGCAGGTG	1
ALDH2	NM_000690	ALDH 2 family (mitochondrial)	12q24.2	1	GTGGGTTGGCTGAGGGT	1
ALDH1A1	NM_000689	ALDH 1 family, member A1	9q21.13	1	TAGCTTCTTCTGAAAGA	3
ALDH18A1	NM_001017423	ALDH 18 family, member A1	10q24.3	3	TAGTCATCTTCAAAAAG	1
ALDH9A1	NM_000696	ALDH 9 family, member A1	1q23.1	1	TTACTCTTTCTCTCTCC	1
Histone modification						
DNMT1	NM_001379	DNA (cytosine-5-)-methyltransferase 1	19p13.2	6	AAGCTGTTGTGTGAGGT	1
DNMT3A	NM_022552	DNA (cytosine-5-)-methyltransferase 3α	2p23	1	CAATAACCCTTTGATTG	1
				1	AGGATGGAGAGAAGTAT	2
HAT1	NM_003642	Histone acetyltransferase 1	2q31.2–q33.1	2	AACAGCTGGAAGAGAGT	1
HDAC10	NM_032019	Histone deacetylase 10	22q13.31	8	CAACCCACGCTCGGTCC	1
HDAC2	NM_001527	Histone deacetylase 2	6q21	4	CTTTATGTGATAGTATT	1
HDAC6	NM_006044	Histone deacetylase 6	Xp11.23	1	GCAAGGTTGCATATGTA	1
HDAC11	NM_024827	Histone deacetylase 11	3p25.1	1	GGATTTGCTGCCCTCTT	1
HDAC7A	NM_015401	Histone deacetylase 7A	12q13.1	7	TTTTTGTAAAAAGGAAG	1
Transcripts expressed in mature blood cells						
ELA2	NM_001972	Elastase 2, neutrophil	19p13.3	26	GGCTGGGGCCTTCTGGG	1
MPO	NM_000250	Myeloperoxidase	17q23.1	42	GCTCCCCTTTTTCTTCC	1
				1	CAAGGCACTGTACTAGG	2
HBB	NM_000518	Hemoglobin, β	11p15.5	2	GCAAGAAAGTGCTCGGT	1

(continued)

Table 1. (Continued)

Symbol	Accession no.	Gene name	Location	Counts	Tag	Position
Transcripts expressed in nonhematopoietic cells						
BTG3	NM_006806	BTG family, member 3	21q21.1–q21.2	1	TAGTTGCAAATAAAAAA	2
SMN1	NM_000344	Survival of motor neuron 1, telomeric	5q13	3	GCTGTTCATTGTACTGT	1
OR4N2	NM_001004723	Olfactory receptor, family 4, subfamily N, member 2	14q11.2	1	AAAAAGGTGTTTAATAA	1
OR7G1	NM_001005192	Olfactory receptor, family 7, subfamily G, member 1	19p13.2	1	CAATTCTCCTGCCTCGG	1
AKR1B1	NM_001628	aldo-keto reductase family 1, member B1 (aldose reductase)	7q35	2	AAGAGTTTTGAAGCTGT	1
AKR1A1	NM_006066	Aldo-keto reductase family 1, member A1 (aldehyde reductase)	1p33–p32	11	GCGTGATCCTGATGAGC	1
Micro-RNA processing						
RNASE3L	NM_013235	Nuclear RNase III Drosha	5p13.3	3	CAAGTGTGGAGTATTTA	1
DICER1	NM_177438	Dicer1, Dcr-1 homolog (*Drosophila*)	14q32.13	1	CTGCAGAAATTTGCAGT	1
DGCR8	NM_022720	DiGeorge syndrome critical region gene 8	22q11.2	5	CTTCAAGGCCGGGGCAG	1

The genes shown in boldface designate transcripts that have not been previously detected in previously published serial analysis of gene expression libraries of cells of a similar phenotype [17, 18]. Tag counts shown are absolute values from a total of 201,106 tags sequenced. Abbreviation: ALDH, aldehyde dehydrogenase.

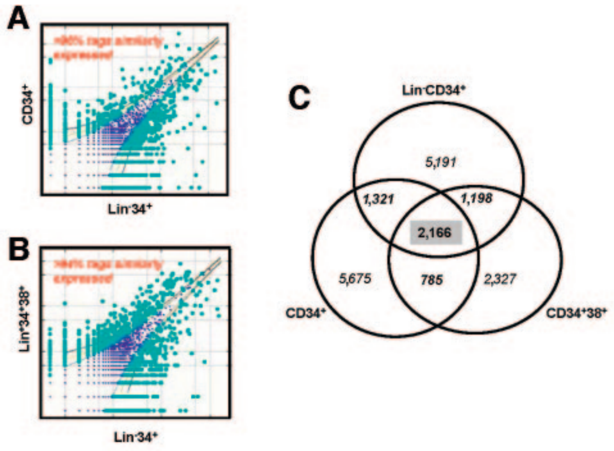


Figure 6. Comparison of polymerase chain reaction-long serial analysis of gene expression (PCR-LongSAGE) library from human lin^{-}CD34^{+} adult bone marrow cells with published data. **(A):** Comparison, using DiscoverySpace software, of the tags present in the PCR-LongSAGE library constructed from the adult human lin^{-}CD34^{+} bone marrow cells in this study and those identified in a published 14-mer serial analysis of gene expression (SAGE) library constructed from a different source of CD34^{+} adult human bone marrow cells [18]. **(B):** Comparison, using DiscoverySpace software, of the tags present in the PCR-LongSAGE library constructed from the adult human lin^{-}CD34^{+} bone marrow cells in this study and those identified in a published 14-mer SAGE library constructed from CD34^{+}CD38^{+} adult human bone marrow cells [20]. **(C):** A Venn diagram showing the intersect of commonly expressed tags in the PCR-LongSAGE, lin^{-}CD34^{+}CD38^{+}, and the CD34^{+} SAGE libraries. To carry out this comparison we converted our LongSAGE tags to short SAGE tags using the DiscoverySpace software. Out of 2,166 tags, 718 could be annotated using the human RefSeq database. These tags and their annotations are listed in supplemental online Table 4.

those inferred from the PCR-LongSAGE tag counts using DiscoverySpace for tag-to-transcript identification (Fig. 5C).

The tag-to-transcript analysis showed that 8,959 tags in the PCR-LongSAGE library mapped to single RefSeq transcripts or multiple variants of a single gene in the RefSeq database. This included transcripts that are known to be expressed in CD34^{+} human bone marrow cells, such as transcripts that encode various transcription factors and cell surface receptors [15–18]. A number of these transcripts have not been found in previously published libraries generated from phenotypically similar cell populations using the original 14-mer SAGE protocol (examples highlighted in Table 1) [17, 19]. Nevertheless, when DiscoverySpace was used to compare all of the tags present in our library with those present in the two related published libraries [18, 20], 96% and 94% similarity values, respectively, were obtained (at a 95% confidence interval; Fig. 6A, 6B). When we compared the nonsingleton tags in the newly constructed lin^{-}CD34^{+} bone marrow library with nonsingleton tags in the other two CD34^{+} human cell libraries, the result showed that 2,166 tags were present in all three (Fig. 6C). The tag-to-transcript mapping of these 2,166 tags yielded 718 RefSeq transcripts (the tag and annotation information are summarized in supplemental online Table 5). The consistent expression of these transcripts in the three CD34^{+} libraries suggests that these genes may play important roles in the maintenance and/or differentiation of human hematopoietic stem/progenitor cells.

Gene Ontology analysis of these 718 RefSeq transcripts showed the presence of cell death-related genes where there was a balance in the positive and negative regulators of cell death. We also observed the presence of several positive regulators of cell growth, reflecting the likelihood that some of the cells in the CD34^{+} subset of human bone marrow are proliferating [20]. In addition, we observed the presence of several transcripts encoding proteasome components and members of the ubiquitination complex (supplemental online Fig. 3). Interestingly, it was recently demonstrated that the proteasomal activity of human hematopoietic progenitor cells prevents their infectability with lentiviral vectors [21].

We also compared our normal adult lin^{-}CD34^{+} human bone marrow cell SAGE library to 287 publicly accessible SAGE libraries prepared from multiple types of human cells (available primarily through the Cancer Genome Anatomy Project at http://cgap.nci.nih.gov, including the two human CD34^{+} cell libraries mentioned above). This more extensive comparison revealed 936 tags that appeared only in our lin^{-}CD34^{+} bone marrow cell library, of which 192 mapped to a single sequence in the human genome and not to any site included in the mammalian genome collection (ftp://ftp.ncbi.nih.gov/repository/MGC/MGC.sequences), RefSeq (ftp://ftp.ncbi.nih.gov/refseq/daily), or Ensembl, version 20. We then estimated the probability of single-base pair errors by combining a library-wide construction error rate and a tag-specific sequencing error probability [22], which indicated that 190 of the 192 tags could be judged to be error-free ($p \leq .05$). Of the 190 tags, 23 mapped to highly conserved regions in mouse, rat, and human genomes and, in the human genome, were located at least 5,000 base pairs away from well-annotated transcripts and were also not present in any human EST database. These 23 novel tags are listed with their chromosomal locations in supplemental online Table 4. Q-RT-PCR was then used to investigate the expression of these 23 novel tags in three cDNA samples prepared from independently from three samples of $\text{lin}^-\text{CD34}^+$ adult human bone marrow cells, including one prepared from the same pool of RNA used for making the PCR-SAGE library.

To assess the possibility of genomic DNA contamination and its contribution to the detection of the unique tag expression, we included a strict negative control in which RNA from each bone marrow sample was used as PCR template (described in Materials and Methods). Q-RT-PCR analyses showed 10 of the 23 tags to be consistently detectable in the cDNA samples examined with no detectable amplification in the negative controls. Four of these 10 novel tags were also observed in nine additional PCR-LongSAGE libraries that we have recently prepared from related sources of primitive human hematopoietic cells (i.e., the $\text{lin}^-\text{CD34}^+\text{CD38}^-\text{CD7}^-\text{CD36}^-\text{CD45RA}^-\text{CD71}^-$ and $\text{lin}^-\text{CD34}^+\text{CD38}^+\text{CD7}^-\text{CD36}^-\text{CD45RA}^-\text{CD71}^-$ subsets of cells in normal adult human bone marrow, umbilical cord blood, G-CSF-mobilized peripheral blood, and human fetal liver; Y.Z. and C.J.E., unpublished data), and 1 of the 10 novel tags was present in two of these nine libraries (supplemental online Table 4).

Discussion

SAGE technology offers a powerful approach to global gene expression profiling of defined cell populations and can serve as an important gene discovery tool. It is therefore particularly attractive for investigations of changes in cellular programs, both normal and aberrant. However, the use of SAGE to interrogate many key events is often precluded because these take place in rare cell types that are inaccessible to SAGE analysis because the amounts of RNA required cannot be obtained. Here, we describe a modified method for preparing amplified cDNA products that enables LongSAGE to be reproducibly applied to samples 10-fold smaller than were previously possible (10^3 cells or less). This modification makes use of a template switching primer containing a rare (*Asc*I) restriction site and a 21-cycle PCR that yields sufficient cDNA product to allow the construction of SAGE libraries from which millions of tags can be derived by direct sequencing. Here, we used the Long-SAGE protocol because of the improved yield of tags obtained from such libraries that can be uniquely mapped to genomic DNA [23].

Currently, many of the methods available to amplify RNA make use of the error-prone T7 RNA polymerase. If applied to material to be used for SAGE, a high frequency of ambiguous or incorrect tags might be expected. Amplification of cDNAs by the PCR method makes use of Titanium Taq polymerase with a TaqStart antibody to provide automatic hot-start PCR, as well as proofreading activity. These latter features maximize reliability by ensuring that the amplified cDNA contains very little product derived from nonspecific cDNA strand amplification or mismatched sequence errors (estimated at 1/50,000 nucleotides). Here, we validated these predictions by a series of experimental and statistical comparisons of the tag or transcript representation in amplified versus nonamplified cDNA preparations and SAGE libraries prepared from these samples. The results demonstrated that PCR-LongSAGE is a reproducible method for performing SAGE analyses on small numbers of cells without significant distortion or loss of transcripts present in the original RNA extract.

The power of this method is illustrated here for the transcriptome analysis of the small fraction of $\text{lin}^-\text{CD34}^+$ cells present in normal adult human bone marrow. These cells are of particular interest because they are highly enriched in hematopoietic stem and progenitor cells [18]. Comparison of the PCR-LongSAGE library obtained from this subset with published (SAGE) libraries prepared from nonamplified cDNA obtained from similar cells showed extensive similarities in tag composition and the presence of many expected transcripts. In addition, our studies underscore the power of the LongSAGE protocol for identifying novel transcripts and transcripts of potential developmental importance because of their restricted but reproducible detection in closely related primitive cell populations. We therefore expect that this method will broaden the application of SAGE to other purified or microdissected subsets of cells and thereby facilitate the investigation of many processes not previously accessible to global gene expression analysis.

Acknowledgments

We are grateful to members of the Stem Cell Research Laboratory and FACS facility in the Terry Fox Laboratory for assistance in the initial processing and FACS isolation of the human cord blood and bone marrow cells used, to Dr. J. Thomson (University of Wisconsin, Madison, WI) for providing the H9 cell RNA extract, and to A. Wanhill and D. Wytrykush for assistance in preparing the manuscript. This work was supported by funds from Genome BC and Genome Canada, the Stem Cell Network, and the Terry Fox Run (as a grant from the National Cancer Institute of Canada). Y.Z. held postdoctoral fellowships from the Stem Cell Network and the Leukemia Research Fund of Canada. A.R. held postdoctoral fellowships from the Canadian Breast Cancer Foundation and the Canadian Institutes of Health Research. D. Kent held Studentships from the Stem Cell Network and the Canadian Institutes of Health Research. M.A.M. and S.J. are Scholars of the Michael Smith Foundation for Health Research, and M.A.M. is a Terry Fox Young Investigator of the National Cancer Institute of Canada. Y.Z. and A.R. contributed equally to this work.

Disclosure of Potential Conflicts of Interest

The authors indicate no potential conflicts of interest.

References

1 Hofman P. DNA microarrays. Nephron Physiol 2005;99:85–89.

2 Velculescu VE, Zhang L, Vogelstein B et al. Serial analysis of gene expression. Science 1995;270:484–487.

3 Datson NA, van der Perk-de Jong J, van den Berg MP et al. MicroSAGE: A modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res 1999;27:1300–1307.

4 Heidenblut AM, Luttges J, Buchholz M et al. aRNA-longSAGE: A new approach to generate SAGE libraries from microdissected cells. Nucleic Acids Res 2004;32:e131.

5 Vilain C, Libert F, Venet D et al. Small amplified RNA-SAGE: An alternative approach to study transcriptome from limiting amount of mRNA. Nucleic Acids Res 2003;31:e24.

6 Peters DG, Kassam AB, Yonas H et al. Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. Nucleic Acids Res 1999;27:e39.

7 Neilson L, Andalibi A, Kang D et al. Molecular phenotype of the human oocyte by PCR-SAGE. Genomics 2000;63:13–24.

8 Van Gelder RN, von Zastrow ME, Yool A et al. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A 1990;87:1663–1667.

9 Zhu YY, Machleder EM, Chenchik A et al. Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. Biotechniques 2001;30:892–897.

10 Hogge DE, Lansdorp PM, Reid D et al. Enhanced detection, maintenance and differentiation of primitive human hematopoietic cells in cultures containing murine fibroblasts engineered to produce human Steel factor, interleukin-3 and granulocyte colony-stimulating factor. Blood 1996;88: 3765–3773.

11 Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 1997;7:986–995.

12 Perlman G, Horan FL. Report on STAT release 5.1 data analysis programs for UNIX and MSDOS. Behav Res Methods Instrum Comput 1986;18:168–176.

13 Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.2. Cladistics 1989;5:164–166.

14 Richards M, Tan SP, Tan JH et al. The transcriptome profile of human embryonic stem cells as defined by SAGE. Stem Cells 2004;22:51–64.

15 Bello-Fernandez C, Matyash M, Strobl H et al. Analysis of myeloid-associated genes in human hematopoietic progenitor cells. Exp Hematol 1997;25:1158–1166.

16 Mao M, Fu G, Wu J-S et al. Identification of genes expressed in human CD34$^+$ hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. Proc Natl Acad Sci U S A 1998;95:8175–8180.

17 Zhou G, Chen J, Lee S et al. The pattern of gene expression in human CD34$^+$ stem/progenitor cells. Proc Natl Acad Sci U S A 2001;98: 13966–13971.

18 Civin CI, Almeida-Porada G, Lee M-J et al. Sustained, retransplantable, multilineage engraftment of highly purified adult human bone marrow stem cells in vivo. Blood 1996;88:4102–4109.

19 Georgantas RW III, Tanadve V, Malehorn M et al. Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. Cancer Res 2004;64: 4434–4441.

20 Ng YY, van Kessel B, Lokhorst HM et al. Gene-expression profiling of CD34+ cells from various hematopoietic stem-cell sources reveals functional differences in stem-cell activity. J Leukoc Biol 2004;75:314–323.

21 Santoni de Sio FR, Cascio P, Zingale A et al. Proteasome activity restricts lentiviral gene transfer into hematopoietic stem cells and is down-regulated by cytokines that enhance transduction. Blood 2006;107: 4257–4265.

22 Siddiqui AS, Khattra J, Delaney A et al. A mouse atlas of gene expression: Large-scale, digital gene expression profiling resource from precisely defined developing C57BL/6J mouse tissue and cells. Proc Natl Acad Sci U S A 2005;102:18485–18490.

23 Saha S, Sparks AB, Rago C et al. Using the transcriptome to annotate the genome. Nat Biotechnol 2002;20:508–512.

See www.StemCells.com for supplemental material available online.

A Modified Polymerase Chain Reaction-Long Serial Analysis of Gene Expression Protocol Identifies Novel Transcripts in Human CD34 + Bone Marrow Cells

Yun Zhao, Afshin Raouf, David Kent, Jaswinder Khattra, Allen Delaney, Angelique Schnerch, Jennifer Asano, Helen McDonald, Christina Chan, Steven Jones, Marco A. Marra and Connie J. Eaves

Stem Cells 2007;25;1681-1689; originally published online Apr 5, 2007;
DOI: 10.1634/stemcells.2006-0794

P061

CYTOGENETIC AND MOLECULAR CHARACTERIZATION OF 979 PATIENTS WITH CHRONIC MYELOPROLIFERATIVE DISEASES AND OF 221 PATIENTS WITH MYELODYSPLASTIC SYNDROMES

U. Bacher[1], T. Haferlach[2], C. Haferlach[2], W. Kern[2], A. R. Zander[1], S. Schnittger[2]
[1]Stem Cell Transplantation, University Hospital of Hamburg-Eppendorf, Hamburg, GERMANY, [2]GmBH, MLL Munich Leukemia Laboratory, Munich, GERMANY

Introduction: Differentiation of *BCR-ABL*-negative chronic myeloproliferative diseases (CMPD) and of myelodysplastic syndromes (MDS) is often difficult due to morphological overlaps. We performed cytogenetic and molecular screening in 979 CMPD, in 221 MDS cases, and in 85 patients with secondary acute myeloid leukemia (s-AML) to evaluate overlaps and disease-specific anomalies.

Methods: 979 patients with CMPD (polycythemia vera (PV): n=225, essential thrombocytosis (ET): n=218, chronic idiopathic myelofibrosis (CIMF): n=57; not classified: n=479) and 221 patients with MDS (refractory anemia (RA): n=3; 5q-: n=16; RARS: n=11, refractory cytopenia with multilineage dysplasia (RCMD): n=20; RA with blast excess (RAEB): n=111; CMML: n=60) were analyzed by cytogenetics and by PCR for the *JAK2*V617F activating mutation. The MDS cases were screened for *FLT3*-length mutations (*FLT3*-LM), *NRAS* mutations, and partial tandem duplications of *MLL* (*MLL*-PTD).

Results: -7 (4/150; 2.7%) and del(7q) (2/150; 1.3%) were exclusively observed in MDS. Del(11q) (2/362; 0.6%), del(12p) (1/362; 0.3%), del(13q) (3/362; 0.8%), and +9 (7/362; 1.9%) occurred only in CMPD - the latter as sole aberration in *JAK2*V617F-mutated cases only. 5q- was more frequent in MDS than in CMPD (20/150; 13.3% *vs.* 3/362; 0.8%). +14, del(20q), and i(17q) occurred in MDS and in CMPD in <2%. *JAK2*V617F was more frequent in the CMPD (795/972; 81.8%) than in MDS (12/116; 10.3%) and in s-AML after a CMPD (16/33; 48.5%) than in s-AML after MDS (2/11;18.2%). In MDS the *FLT3*-LM were found in 4/152 (2.6%), *NRAS* in 8/106 (7.5%), and *MLL*-PTD in 8/153 (5.2%).

Conclusions: The overlaps of the *JAK2*V617F-mutation and 5q-, +8, del(20q), and i(17q) in both disorders suggest common leukemogeneic pathways in part of MDS and CMPD cases. The occurrence of *FLT3*-LM, *NRAS*, *MLL*-PTD, and chromosome 7 anomalies in MDS and of trisomy 9 in the CMPD especially in *JAK2*V617Fmut cases can be helpful in the differentiation of both disorders.

U. Bacher, None.

P062

COMPARATIVE TRANSCRIPTOME ANALYSIS OF NORMAL AND CHRONIC MYELOID LEUKEMIA STEM CELLS

Y. Zhao[1], A. Delaney[2], M. A. Marra[2], A. C. Eaves[1], C. J. Eaves[1]
[1]Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC, CANADA, [2]Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, CANADA

Chronic myeloid leukemia (CML) arises from a hematopoietic stem cell that acquires a BCR/ABL fusion gene. To obtain new insights into the molecular perturbations characteristic of CML stem cells, we generated serial analysis of gene expression (SAGE) libraries (~200,000 tags/library) from extracts of highly purified $lin^-CD34^+CD45RA^-CD36^-CD71^-CD7^-CD38^+$ and $lin^-CD34^+CD45RA^-CD36^-CD71^-CD7^-CD38^-$ normal bone marrow (BM) and G-CSF-mobilized peripheral blood (G-mPB) cells as well as cells from 3 chronic phase CML patients with predominantly Ph+/BCR-ABL+ cells in both subsets. Comparison of the tags present in pooled CML and pooled normal BM and G-mPB libraries revealed many candidate differentially expressed genes. Real-time RT-PCR analysis of lin^-CD34^+ cells from 14 chronic phase CML patients and 3 normal BMs confirmed the differential expression of 13 candidates identified by SAGE (changes ranging from 3 fold lower to ~80 fold higher in the CML cells, $p<0.05$). The altered levels of expression of 5 of these genes were highly correlated with the relative levels of BCR-ABL transcripts in the same cells ($r \geq 0.6$). Moreover, 5 of these 13 genes were differentially expressed in $CD34^+$ cord blood cells analyzed 3 days post-transduction with a BCR/ABL-IRES-GFP lentivirus by comparison to $CD34^+$ cells transduced with an empty GFP vector (n=2). In addition, we identified 65 unique tags in the 38^- subset of CML cells from a comparison of the meta-CML 38^- cell library with the normal meta libraries (for both 38^- and 38^+ cells), the meta 38^+ cell CML libraries and most publicly accessible SAGE data. 32 of these unique tags were located within conserved genomic regions and >2kb away from known transcripts, and 3 were confirmed to represent novel transcripts using a PCR approach. These results illustrate the potential of SAGE to reveal novel as well as known components in the transcriptomes of rare normal and cancer stem cell populations.

Y. Zhao, None.

P063

MODULATION OF THE LEUKEMIC STEM CELL ACTIVITY IN MN1-INDUCED LEUKEMIAS BY NUP98HOXD13

M. Heuser, B. Argiropoulos, R. K. Humphries
Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC, CANADA

Overexpression of HOX genes characterize the large group of AML patients with NPM1 mutations whereas overexpression of MN1 is highly correlated with wildtype-NPM1 AML (Verhaak et al. 2005). To test whether HOX and MN1 pathways are mutually exclusive we coexpressed NUP98HOXD13 (ND13), a myeloid oncogene, in an MN1-mediated leukemia model. Mice transplanted with 1x10E6 cells overexpressing either MN1 or MN1+ND13 died after a median latency of 35 and 40 days, respectively (P=0.26). However, in limiting-dilution analysis of MN1 expressing cells the frequency of leukemic-stem cells (LSC) was one in 5465 (0.018 percent), whereas in contrast to our hypothesis it was 33-fold increased in MN1+ND13 coexpressing cells. In addition, the disease latencies at limiting dilution differed significantly (MN1: 82 vs. MN1+ND13: 44 days, P=0.009), demonstrating that the addition of ND13 to MN1 enhanced the potency of the individual LSC besides its frequency. To better understand non-redundant and redundant downstream pathways expression of 13 genes reportedly involved in stem-cell regulation was quantified in MN1-only, ND13-only, or MN1+ND13 cells and compared to their expression in normal bone marrow cells. We found that Bmi-1 and mel18 were exclusively upregluated by MN1, HoxA7 exclusively upregulated by ND13, and HoxA9 and Gata-2 upregulated by both genes. Jak2 was significantly downregulated by MN1 only, whereas Notch-1, Lnk, c-mpl, Cxcl-12, and Vcam-1 were downregulated by both genes. Expression of SOCS-1 and rae28 was not affected. We provide a model that allows us to modulate the activity of LSCs and to explore the regulation and signaling of LSCs. Functional and gene expression data point to non-redundant signaling pathways in MN1 and ND13-transduced cells. Although MN1 and HOX-gene expression are inversely correlated in human AML, they synergize in our model. This

Blood (ASH Annual Meeting Abstracts) 2007 110: Abstract 35

Oral Session

Chronic Myeloid Leukemia: Mechanisms of Disease and Genomic Instability

Chronic Myeloid Leukemia - Biology and Pathophysiology excluding Therapy

Comparative Transcriptome Analysis of Different Subsets of CD34^{+} Normal and Chronic Myeloid Leukemia Cells Identifies Novel Perturbations in the CML Stem Cell Population.

Yun Zhao, PhD[1,*], Allen Delaney, PhD[2,*], Marco A. Marra, PhD[2,*], Xiaoyan Jiang, MD, PhD[1], Allen C. Eaves, MD, PhD[1] and Connie J. Eaves, PhD[1]

[1] Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC, Canada and [2] Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada.

Abstract

Chronic myeloid leukemia (CML) arises from a hematopoietic stem cell that acquires a *BCR-ABL* fusion gene. During the chronic phase of the disease, this cell produces an expanding multi-lineage clone that usually comes to dominate all terminal stages of myelopoiesis. However, eventually, further mutations are acquired which cause progression to a rapidly fatal acute leukemia. To obtain new insights into the molecular perturbations that cause CML stem cells to initiate chronic phase disease, we generated Long Serial Analysis of Gene Expression (LongSAGE) libraries (~200,000 tags/library) from extracts of highly purified lin^{-}CD34^{+}CD45RA^{-}CD36^{-}CD71^{-}CD7^{-}CD38^{+} and lin^{-}CD34^{+}CD45RA^{-}CD36^{-}CD71^{-}CD7^{-}CD38^{-} normal bone marrow (BM) and G-CSF-mobilized peripheral blood (G-mPB) cells as well as cells from 3 chronic phase CML patients with predominantly Ph^{+}/BCR-ABL^{+} cells in both of these very primitive cell subsets. Long term culture-initiating cell (LTC-IC) and direct colony-forming cell (CFC) assays performed on an aliquot of each of these cell populations showed the frequency of LTC-IC were 4 to 130-fold higher in the 34^{+}38^{-} cells than in the matching 34^{+}38^{+} cells with the opposite trend for the CFCs. Comparison of the tags present in the pooled CML and pooled normal BM and G-mPB libraries revealed many differentially expressed genes. Real-time RT-PCR analysis of lin^{-}CD34^{+} cells from 14 chronic phase CML patients and 3 normal BMs confirmed the differential expression of 14 candidate transcripts identified by SAGE (changes ranging from 3-fold lower to 80-fold higher in the CML cells, $p<0.05$). The altered levels of expression of 5 of these genes (i.e., *beta-catenin*, *MLLT3, IL1R1, LY6E* and GAS2) were highly correlated with the relative levels of *BCR-ABL* transcripts in the same cells ($r \geq 0.6$). 5 of the 14 genes (*IL1R1*, *vWF, SOX4*,

SELL and *RHOB*) were found to be differentially expressed in the 3-day post-transduction progeny of CD34$^+$ cord blood cells exposed to a *BCR/ABL-IRES-GFP* vs a control *GFP*-lentivirus preparation (n=2). 3 (*GAS2, DUSP1 and TP53BP2*) were upregulated (5 to 11-fold) in imatinib-treated K562 cells (as compared to untreated K562 cells) but their expression remained unchanged in similarly treated KG1 cells (a primitive *BCR-ABL*-negative human AML cell line) providing further evidence that their deregulated expression is secondary to the kinase activity of p210$^{\text{BCR-ABL}}$. In addition, from a comparison of the meta-library for the 34$^+$38$^-$ CML cells with the meta-libraries for both the normal 34$^+$38$^-$ and 34$^+$38$^+$ cells, the meta-library for the 34$^+$38$^+$ CML cells and most publicly accessible SAGE data, we were able to identify 65 novel tags in the 34$^+$38$^-$ CML cells. 32 of these unique tags were located within conserved genomic regions and >2 kb away from known transcripts, and of these 32, 3 were confirmed to represent novel transcripts using a PCR approach. These results illustrate the potential of SAGE to reveal novel as well as known components in the transcriptomes of rare normal and cancer stem cell populations. Investigation of their roles in primitive human cells transduced with *BCR-ABL* and *BCR-ABL*$^+$ cell lines indicates the utility of these models for further delineation of the complex effects of *BCR-ABL* expression in chronic phase CML stem cells.

Footnotes

Disclosure: No relevant conflicts of interest to declare.

Blood (ASH Annual Meeting Abstracts) 2008 112: Abstract 193

Oral Session

Chronic Myeloid Leukemia - Stem Cell Biology and Eradication

Differentially Expressed and Novel Transcripts in Highly Purified Chronic Phase CML Stem Cells

Yun Zhao[1,*], Allen Delaney[2,*], Afshin Raouf[1,*], Kamini Raghuram[1,*], Haiyan I Li[2,*], Angelique Schnerch[2,*], Xiaoyan Jiang, MD, PhD[1], Allen C Eaves[1,*], Marco A Marra, PhD[2] and Connie J. Eaves, PhD[1]

[1] Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, BC, Canada, [2] Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

Abstract

The chronic phase of CML is sustained by rare BCR-ABL+ stem cells. These cells share many properties with normal pluripotent hematopoietic stem cells, but also differ in critical ways that alter their growth, drug responsiveness and genome stability. Understanding the molecular mechanisms underlying the biological differences between normal and CML stem cells is key to the development of more effective CML therapies. To obtain new insights into these mechanisms, we generated Long Serial Analysis of Gene Expression (SAGE) libraries from paired isolates of highly purified lin-CD34+CD45RA-CD36- CD71-CD7-CD38+ and lin-CD34+CD45RA-CD36-CD71-CD7-CD38- cells from 3 chronic phase CML patients (all with predominantly Ph$^+$/BCR-ABL+ cells in both subsets) and from 3 control samples: a pool of 10 normal bone marrows (BMs), a single normal BM and a pool of G-CSF-mobilized blood cells from 9 donors. In vitro bioassays showed the CD34+CD38+ cells were enriched in CFCs (CML: 3–20% pure; normal: 4–19% pure) and the CD34+CD38- cells were enriched in LTC-ICs (CML: 0.2–26% pure; normal: 12–52% pure). Each of the 12 libraries was then sequenced to a depth of ~200,000 tags and tags from libraries prepared from like phenotypes were compared between genotypes using DiscoverySpace software and hierarchical clustering. 1687 (355 with clustering) and 1258 (316 with clustering) transcripts were thus identified as differentially expressed in the CML vs control CD34+CD38– and CD34+CD38+ subsets, respectively. 266 of these transcripts (11 with clustering) were differentially expressed in both subsets. The differential expression of 5 genes (*GAS2*, *IGF2BP2*, *IL1R1*, *DUSP1* & *SELL*) was confirmed by real-time PCR analysis of lin-CD34+ cells isolated from an additional 5 normal BMs and 11 CMLs, and lin-CD34+CD38– cells from an additional 2 normal BMs and 2 CMLs (with dominant Ph$^+$ cells). *GAS2* and *IL1R1* transcript levels were correlated with *BCR-ABL* transcript levels in both primitive subsets, and predicted differences in expression of *IL1R1* and *SELL* were apparent within 3 days in CD34+ cord blood cells transduced with a lenti-*BCR-ABL*-IRES-GFP vs a control lenti-GFP vector (n=3). These findings support a direct role of BCR-ABL in perturbing the expression of these 3 genes. Further comparison of the meta CD34+CD38– and CD34+CD38+ CML cell libraries with most publicly accessible SAGE data revealed 69 novel tags in the CD34+ CML cells that correspond to unique but conserved genomic sequences. Nine of these were recovered by 5'- and 3'- RACE applied to cDNAs pooled from several human leukemic cell lines. These results illustrate the power of SAGE to reveal key components of the transcriptomes of rare human CML stem cell populations including transcripts of genes not previously known to exist. Continuing investigation of their biological roles in primary CML cells and primitive *BCR-ABL*-transduced human cells offer important strategies for delineating their potential as therapeutic targets.

Footnotes

Corresponding author

Disclosures: No relevant conflicts of interest to declare.